PREDICTIVE MODEL FRAMEWORK FOR POWER HOLDING COMPANY OF NIGERIA

ISSN: 2141 – 3290
www.wojast.com

IGODAN, E. C. & *UKAOHA, K. C.
Department of Computer Science,
University of Benin, Benin City, Nigeria.
**e-mail: kingsley.ukaoha@uniben.edu*
**corresponding author*

ABSTRACT

In recent years, data mining has attracted a great deal of attention due to the availability of huge data and the urgent need for turning such data into useful information and knowledge. The Power Holding Company of Nigeria (PHCN) has been the sole electricity company in Nigeria. It has been faced with enormous amount of data on a daily bases even before the introduction of e-prepaid electricity metres. The lack of ability to determine and detect frauds from within the densely populated residential houses has bedevilled the PHCN. Although, "We are living in the information age" is a popular saying, but the truth is that we are actually living in the data age. This has led to uninformed decision making. Consequently, decisions are subjectively made. This research work was aimed at using exploratory data analysis and data mining technique in building a predictive model for decision support in predicting customer's behaviour. The data set collected were cleaned, aggregated, segmented and normalized using the min-max technique, the Euclidean distance and correlation coefficient calculated, and k-means clustering technique, a non-hierarchical method, was adopted. The data set were visualized using data visualization tool to present the results for knowledge discovery. The system was implemented using XAMP software on a Windows XP OS platform.

INTRODUCTION

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large data repositories become "data tombs" - data archives that are seldom visited. Consequently, important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. In addition, consider expert system technologies, which typically rely on users or domain experts to *manually* input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific, medical research and according to Kraft *et al.* (2002) thorough analysis of available data on a given problem can lead to more efficient decision making. The widening gap between data and information calls for a systematic development of *data mining tools* that will turn data tombs into "golden nuggets" of knowledge. The generation of information and knowledge calls for data organised into a useful form (Kraft *et al.* 2002). The main idea of this study was to develop a predictive model for the Power Holding Company of Nigeria to use in predicting the behaviours of consumers living in the densely residential areas in detecting frauds using data mining concepts.

Numerous works in literatures using data mining and data exploratory techniques on different

areas of application have motivated this work.

In Bastos *et al.* (2005) a maintenance behaviour-based prediction system was developed using data mining concepts. The idea was to predict the possibility of breakdowns in the industrial manufacturing with bigger accuracy for increased system reliability. In the study, an organizational architecture that integrates data produced from factories on their activities of reactive, predictive and preventive maintenance was proposed. They posited that the system will help enterprises to collect, extract and create knowledge in a way that enterprise will predict with more accuracy the moment to realise maintenance actions and thus improve the productivity of manufacturing process. Adams *et al.* (2011) used univariate time series models to forecast electricity generation in Nigeria. The paper examined the appropriate model that fits the aggregate electricity generated in Nigeria between 1970 and 2009. It was discovered that the Box-Jenkins Autoregressive Integrated Moving Average model (ARIMA) (3,2,1) is the most suitable model for the series with the Normalized Bayesian Information Criteria (BIC) of 13.906, stationary $R^2 = 0.69$ and Maximum likelihood estimate of 411.55 and the Ljung-Box test ($Q_{14} = 6.404$ and $p>.10$) was also estimated. The ARIMA model revealed that the average electricity generation in Nigeria will reduce further, with only a slight increment forecasted in year 2011. In Dhanda (2011) an innovative utility sentient approach for the mining of interesting association patterns from transaction database to illustrate the limitations of apriori algorithm was proposed. The frequent patterns were first discovered from the database and then mined. In the study, the approach extracts novel interesting association patterns with emphasis on significance, quantity, profit and confidence; and a comparative analysis was also presented to illustrate the effectiveness of the approach. Shekar and Srinivas (2008) classified sample data according to age, gender and type of refractive errors in school-going children applying the data mining technique decision tree that was constructed using ID3 Algorithm. Venkateswari and Suresh (2011) carried out a survey of an association rule mining and its various applications in e-commerce environment. In Delen *et al.* (2004) a comparative study of multiple prediction models for breast cancer survivability using two popular data mining algorithms-artificial neural network and decision trees, along with a most commonly used statistical method-logistic regression on a large dataset was carried out. In the study, 10-fold cross-validation was used to measure the unbiased estimate of the three prediction models for performance. In Kraft *et al.* (2002) the length of stay of a subset of the total population specifically those with Spinal Cord Injury (SCI) was predicted using nursing diagnosis and neural networks. The authors also suggested that SCI patients do not present large numbers, because they are outliers in the healthcare system due to extended hospital stays and high costs for treatment.

METHODOLOGY

The extraction of significant patterns from the data set gathered from the PHCN is discussed in this section. The data warehouse contains data of PHCN customer's for the period of three years. The data warehouse is preprocessed to make the mining process more efficient. Then the min-max technique was applied on the preprocessed data warehouse so as to normalize the values. Thereafter the data was then clustered using the K-means clustering algorithm with K=2. Thereafter, the Euclidean distance is calculated to determine the search space of the values. This results in two clusters which enable the prediction of customer's behaviour.

Data Pre-processing

The data set was cleaned, aggregated, segmented and normalized applying the Min-max technique. The actions involve the removal of duplicate of records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. In order for making the data appropriate for mining process, it needs to be transformed i.e. changed into a more appropriate form. In our approach, the data warehouse

was refined by removing duplicate records, normalized, and transformed to a form appropriate for clustering. Normalization is a process where numeric columns are transformed using a mathematical function to a new range. The mean and min-max approach shown in equation 1.1 was adopted as the normalization while the Euclidean distance given in equation 1.2.

$$value' = \frac{value - OriginalMin}{OriginalMax - OriginalMin}(NewMax - NewMin) + NewMin \quad 1.1.$$

$$d = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad 1.2.$$

K-Means Clustering Algorithm

There are three popular methods for grouping data set: clustering, associative rule, and decision tree (Myatt, 2007). The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a general feature, frequently the proximity with regard to some defined distance measure is known as clustering. The clustering problem has been addressed in numerous contexts besides being proven beneficial in many applications. The K-Means Clustering approach, a non-hierarchical method for grouping unsupervised data set, was adopted in this study.

The K-Means algorithm is given as:

a. Choose a value of k
b. Select k objects in an arbitrary fashion. Use these as the initial set of k centroids
c. Assign each of the objects to the cluster for which it is nearest to the centroid.
d. Recalculate the centroids of the k clusters.
e. Repeat steps 3 and 4 until the centroids no longer move.

Scatter plot Representation

The graphical representations of the results of the data set are depicted using the scatterplot diagram. The variables were then plotted on a scatter plot graph to visualize the relationships. The relationship of the clusters categorized the customers' behaviours into two. Customers with high and low rate credits purchase. The various graphs from the three years of 20 consumers are shown in Figures 1 to 6.

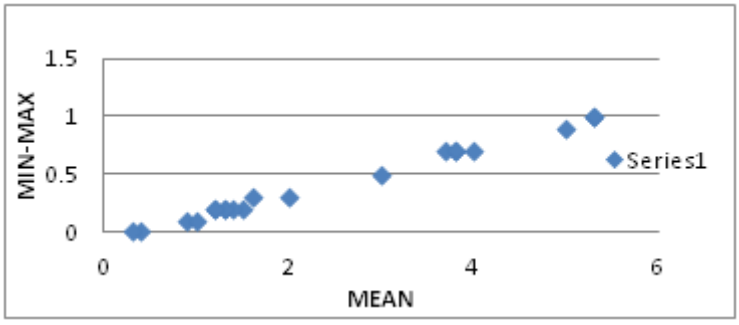

Figure 1: Clusters of Customers' Monthly Vending Dates – Year1

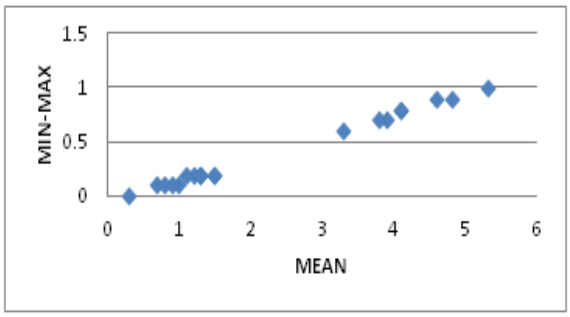

Figure 2: Clusters of Customers' Monthly Vending Dates – Year 2

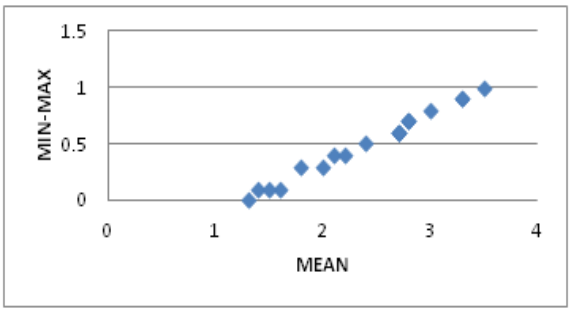

Figure 3: Clusters of Customers' Monthly Vending Dates – Year 3

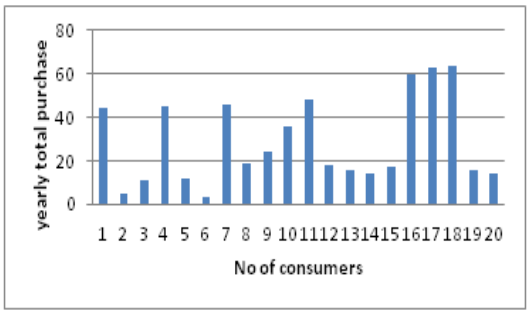

Figure 4: Rate of yearly purchases (for year1)

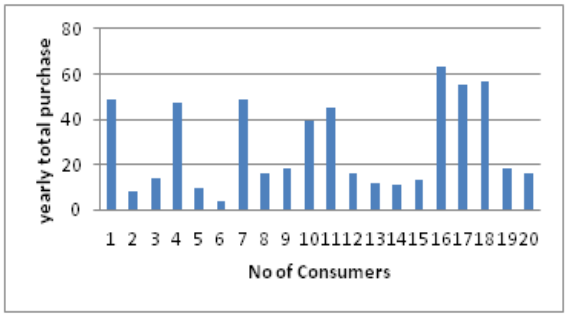

Figure 5: Rate of yearly purchases (for year 2)

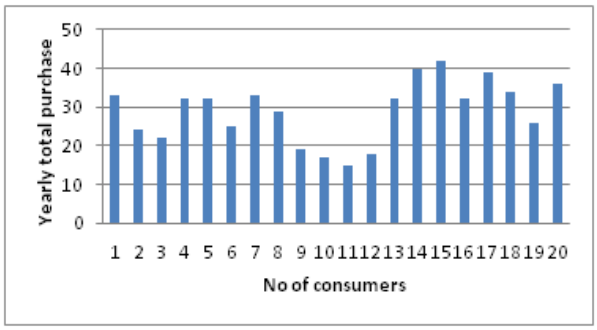

Figure 6: Rate of yearly purchases (for year 3)

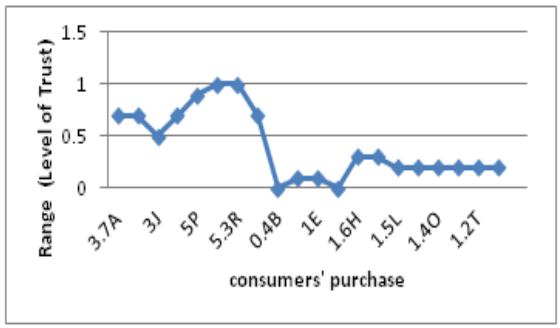

Figure 7: Range - Level of Trust (for year 1).

RESULTS

The data set were separated into three years to ascertain the accuracy and efficiency of the approach used. Using the training set in the third year (year 3) the set was divided into 2 clusters. The cluster was divided into 2 because we either have those consumers that are faithful in their purchase of cards or not faithful. From the range of values of the min-max of 0 to 1, cluster 1falls in between the range of 0 to 0.49 while cluster 2 falls within the range of 0.50 to 1. The set in cluster 1 shows that these set of consumers are to be checked for fraud because of their low purchases while cluster 2 shows high purchases. The lowest and highest values in Cluster 1 are 15 and 26 while cluster 2 has 29 and 42 respectively. Eight of the observations were grouped into cluster 1 while 12 grouped in cluster 2 respectively. This is depicted in the above diagrams as shown in Figures 3 and 6 respectively. Applying the same approach to the data set for year 2, we discovered that cluster 1 falls within the range of 0.0 to 0.49 while cluster 2 falls within 0.50 to 1.2 respectively. The lowest and highest values for clusters 1 and 2 are 8, 18 and 39, 63 respectively. Twelve of the observations were grouped into

cluster 1 while the rest 8 were grouped in cluster 2. In year 1, the lowest and highest values for clusters 1 and 2 are 36, 64 and 3, 24 respectively. While there are 12 observations grouped as cluster 2, 8 were grouped as cluster1. The graphical representation of each year is shown in Figures 7 to 9

Figure 8: Range - Level of Trust (for year 2).

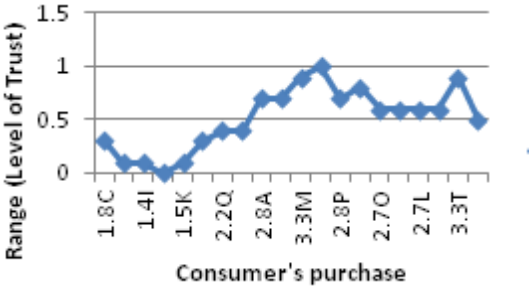

Figure 9: Range - Level of Trust (for year 3).

REFERENCE

Adams, O.S., Akano, O.R., and Asemota, O.J (2011).“Forecasting Electricity Generation in Nigeria using Univariate Time Series Models.” *European Journal of Scientific Research* 1450-216X, 58 (1), pp 30-37.

Bastos, P., Lopes, R., Pires, L., Pedrosa, T. (2009). “Maintenance Behaviour-Based Prediction System Using Data Mining.”*Proceeding of the 2009 IEEE IEEM*. 2487-2491.

Delen, D., Walker, G., and Kadam, A. (2004). “Predictive breast cancer survivability: a comparison of three data mining methods.” *Artificial Intelligence in Medicine* Doi:10.1016/j.artmed. 7(2).

Dhanda, M (2011).“An Approach to Extract Efficient Frequent Patterns from Transactional Database.”*International Journal of Engineering Science and Technology*. ISSN: 0975-5462. 3(7), pp 5652-5658.

Kraft, R.M, Desouza, K.C., and Androwich, I. (2002). “Data Mining in Healthcare Information Systems: Case Study of a Veterans Administration Spinal Cord Injury Population.” *Proceedings of the 36th Hawaii International Conference on System Sciences* (HICSS’03).

Myatt, J.G. (2007). “Making Sense of Data: *A Practical Guide to Exploratory Data Analysis and Data Mining*.” ISBN-13: 978-0-470-07471-8. John Wiley & Sons, Inc. pp 156.

Shekar, V.D and Srinivas, S. (2008). “Clinical Data Mining – An Approach for Identification of Refractive Errors.”*Proceedings of the International Multi Conference of Engineers and Computer Scientists*, (I).

Venkateswari, S and Suresh, M.R. (2011). “Association Rule Mining in e-Commerce: A Survey.” *International Journal of Engineering Science and Technology*, 3 (4), pp. 3086-3089.