Customer Churn Analysis In Banking Sector Using Data Mining Techniques

A. O. Oyeniyi & A.B. Adeyemo
Department of Computer Science,
University of Ibadan
Ibadan, Nigeria
aoyeniyi1@gmail.com, sesanadeyemo@gmail.com

ABSTRACT

Customer churn has become a major problem within a customer centred banking industry and banks have always tried to track customer interaction with the company, in order to detect early warning signs in customer’s behaviour such as reduced transactions, account status dormancy and take steps to prevent churn. This paper presents a data mining model that can be used to predict which customers are most likely to churn (or switch banks). The study used real-life customer records provided by a major Nigerian bank. The raw data was cleaned, pre-processed and then analysed using WEKA, a data mining software tool for knowledge analysis. Simple K-Means was used for the clustering phase while a rule-based algorithm, JRip was used for the rule generation phase. The results obtained showed that the methods used can determine patterns in customer behaviours and help banks to identify likely churners and hence develop customer retention modalities.

Keywords: Customer, banking, data mining, churn analysis, WEKA, retention models & K-means.

African Journal of Computing & ICT Reference Format:

1. INTRODUCTION

The regulatory framework within which financial institutions and insurance firms operate require their interaction with customers to be tracked, recorded, stored in Customer Relationship Management (CRM) databases, and then data mined the information in a way that increases customer relations, average revenue per unit (ARPU) and decrease churn rate. According to [23], churn has an equal or greater impact on Customer Lifetime Value (CLTV) when compared to one of the most regarded Key Performance Indicator (KPI’s) such as Average Revenue Per User (ARPU). As one of the biggest destructors of enterprise value, it has become one of the top issues for the banking industry. Customers churn prediction is aimed at determining customers who are at risk of leaving, and whether such customers are worth retaining.

Churn or customer attrition is a term adopted to define the movement of customers from one provider to another [15], and it is also regarded as the annual turnover of the market base, while recognizing the fact that it cost five (5) times more to acquire new customers than to retain existing customers’ database, as companies often spend fortune on advertisement to acquire new customers [17]. Therefore banks now need to shift their attention from customer acquisition to customer retention, provide accurate churn prediction models, and effective churn prevention strategies as added customer retention solutions to preventing churn [24]. And as [18] also observed better products, convenience and lower fees are not enough to prevent customer churn.

The banking industry needs to intensify campaign to deliver a more efficient, customer focused and innovative offerings to reconnect with their customers. The problem of churn analysis is not peculiar to the banking industry. Churning is an important problem that has been studied across several areas of interest [16], such as mobile and telephony [2]; [3]; [17], insurance [27], and healthcare [9], [4]. Other sectors where the customer churn problem has been analysed includes online social network churn analysis [20]; [1], and the retail banking industries [8]; [12]; [18].

1.1 Data Mining

Data mining is an important component of every CRM framework that facilitates analysis of business problems, prepare data requirements, and build, validate and evaluate models for business problems [32]. The data mining process and algorithms enable firms to search, discover hidden patterns and correlations among data, and to extract relevant knowledge buried in corporate data warehouses, in order to gain broader understanding of business. Data mining uses sophisticated statistical data search algorithms to find, discover hidden patterns and relationships, and extracts knowledge buried in corporate data warehouses, or information that visitors have dropped about their experience, most of which can lead to improvements in the understanding and use of the data in order to detect significant patterns and rules underlying consumer’s behaviours.
Data mining involves four tasks; classification, clustering, regression and association learning; which are classified into two types of data mining; verification-oriented (where the system verifies the user’s hypothesis) and discovery-oriented (where the system finds new rules and patterns autonomously). Data mining process compliment other data analysis techniques such as statistics, on-line analytical processing (OLAP), spreadsheets, and basic data access.

1.2 Data Mining Techniques

Generally, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects) [14]. Decision tree is a symbolic learning technique that organizes information extracted from a training dataset in a hierarchical structure composed of nodes and ramifications. The tree-like output of decision tree makes it easy to understand and interpret, making it the mostly widely used data mining algorithms in many domain such as supplier selection and email user churn analysis [13]. It is capable of building models based on numerical and categorical datasets. Decision tree is also used for classification patterns or piecewise functions. Cluster analysis is an approach by which a set of instances (without predefined class attribute) is grouped into several clusters based on information found in the data that describes the objects and their relationships [30]. A cluster uses a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in another cluster. While in classification the classes are defined prior to building the model, cluster analysis divides the data based on their similarities. There are different types of clustering from different point of view. The most common types divide them all into two types, partitional and hierarchical methods. Partitional clustering is a simple division of a set of data objects into non-overlapping segments such that each data object is in exactly one segment and if we permit clusters to have sub-clusters then we have hierarchical clustering.

1.3 Model Performance Evaluation

The performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the costs of different errors vary remarkably [6]. All classifier’s performance evaluation involves certain level of trade-off between true positive (TP) rate and true negative (TN) rate; and the same applies for recall and precision. Precision, Recall and F-Measure are commonly used in information retrieval as performance measure [5].

1.4 Related Works

Table 1 presents a listing of some of the techniques used in some of the related work on churn prediction in the banking domain.

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Data Mining Techniques</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naveen, N., Ravi, V., &amp; Kumar, D. A. [21]</td>
<td>fuzzyARTMAP (Neural Network Architecture)</td>
<td>Banking (Credit Cards)</td>
</tr>
<tr>
<td>Zhao, Y., Li, B., Li, X., Liu, W., &amp; Ren [34]</td>
<td>Support Vector Machine</td>
<td>Banking</td>
</tr>
<tr>
<td>Farquad, M. A., Ravi, V., &amp; Raju, S. B. [11]</td>
<td>Support Vector Machine (SVM), Naïve Bayes Tree (NBTree), SVM+NBTree Hybrid</td>
<td>Banking (Credit Cards)</td>
</tr>
<tr>
<td>Teemu, M., Ahola, J., &amp; Nousiainen, S. [28]</td>
<td>Logistic Regression</td>
<td>Retail Banking</td>
</tr>
<tr>
<td>Nie, G., Wang, G., Zhang, P., Tian, Y., &amp; Shi, Y. [25]</td>
<td>Logistic Regression</td>
<td>Banking (Credit Cards)</td>
</tr>
</tbody>
</table>
2. MATERIALS AND METHODS

The methodology used for this study is illustrated diagrammatically in Figure 1.

![Research methodology flowchart](image)

**Figure 1: Research methodology flowchart**

2.1 Data Acquisition

The dataset used for this study for customer churn prediction was acquired from a major Nigerian bank. The raw data was extracted from the bank's customer relationship management database and transactional data warehouse which contained more than 1,048,576 customer records described with over 11 attributes. Attributes such as customer name, account number, record start, record closed descriptor that do not affect the customer churn prediction, and or tend to violate the privacy and confidentiality status of the customer records were identified and removed. Also, attributes with a lot of missing values were removed due to the fact that it was difficult recreating values that can fit in for omitted attributes like date of birth (recorded as age), customer type and account type.

The input attributes descriptors used were those for:

1. Customer demographics is the geographic and population data of a given customer or, information about a group living in a particular area;
2. Account level is the billing system including charges; and
3. Customer behaviour is any behaviour related to a customer's bank account.
The customer records had different account types like individual account and corporate account types. The corporate account type had a lot of missing values that concerns customer demographics such as date of birth, and gender. According to [22] customer demographics have been widely used to differential how a segment of customers differs from one another. In determining customer churn or switching in banking, customer demographics such as age, and job type were shown to have an effect on customers switching banks. Other studies proposed inclusion of additional demographics characteristics such as gender, race and occupation as they have great impact on customers switching behaviour in the banking industry. Hence the corporate account types were not used in the study.

2.2 Data Preparation

Data preparation tasks consider transforming acquired datasets to remove noise, inconsistencies, incoherence, bias and redundancies. The data preparation tasks includes table, record, and attribute selection as well as transformation and elimination of data for modelling, that can be performed multiple times, in no prescribed order. The preliminary diagnosis is conducted on the datasets to gain an insight into their properties by scaling or standardizing the data to reduce the level of dispersion between the variables in the datasets [19]. The dataset was pre-processed in Waikato Environment for Knowledge Analysis (WEKA) to clean, transform and establish relationship between the input variables and the output variables.

2.3 WEKA Machine Learning Workbench

The Waikato Environment for Knowledge Analysis otherwise known as WEKA, is a collection of machine learning algorithms for data mining tasks, which can either be applied directly to a desired dataset or invoked from within a java code. The WEKA machine learning workbench was developed by Machine Learning Group at the University of Waikato New Zealand and distributed under GNU General Public license. Embedded within the WEKA workbench are variants of machine learning tools such as data pre-processing and visualisation tools, classification and regression techniques, clustering and association rule mining techniques, which are well suited for developing new models and machine learning schemes. The WEKA machine learning workbench has a unique file format known as Attribute-Relation File Format (.arff) for converting and pre-processing datasets for analysis and evaluation. It also accept datasets saved in a command-separated value file format (.csv), using Load Converter function to convert (.csv) to (.arff), and can connect and load datasets into WEKA from database and website (url). Figure 2 shows part of the bank dataset in an attribute-relation file format (.arff).
3.0 RESULTS AND DISCUSSION

In this study churn prediction was modeled using K-Means clustering techniques and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) also known as JRip algorithm in WEKA. The prepared data was used to generate clusters with similar attributes and to generate rule sets using JRip algorithm. The raw banking datasets acquired consists of 1,048,576 customer’s records, with 11 attributes descriptors. After rigorous data cleaning and transformation, customer’s records considered for final analysis consists of 4958 customers banking records with 8 attributes descriptors of which 500 customer records with four attribute descriptors were used for the study.

For the classifier that was used the datasets was divided into training, testing and cross validation datasets using percentage split and k-fold cross-validation to avoid training and testing on the same data that could leads to false result. The data was split as 66% for training and testing with the remaining 34%. 10-fold stratified cross-validation was used. The Sample data in Excel (.CSV) Format is presented in Figure 3.

Figure 2: Bank dataset in an attribute-relation file format
3.2 Cluster Analysis

Due to the nature of the segment of the banking sector which was used which was non-contract based, it is important to give an appropriate definition of churn prior to building the prediction model. In almost all studies reviewed, bank customers are those customers who had relationship with the bank. Consequently, “churn” in such condition could be defined as the terminating of contract from the customer’s side or not reactivating an account(s) after going into dormancy. However, in the banks there is no contract between the bank and its customers. The customer through marketing or self-will can simply decide to open either individual or corporate account(s) type and automatically become a customer. On the other hand, at any time, customers can stop operating their accounts with the bank, and become a churner without leaving immediate trace. This implies that churn in such cases happens with no tracking point such as closing of account or inactive account this makes it difficult to recognize churners. For example if a customer database consisting of a number of customers with different transactional activities, some of which perform daily/weekly activity on their account either by walking into the banking hall or using the online banking system platform; and some who do not perform any transactional activity on their account, is considered. And if a churner is defined as “a person who has not used his/her account for 3 months”, then a considerable part of the customers who use their account occasionally, for instance every 2 months, would be mistakenly considered as a churner. If a longer time span is used for the prediction period and a churner is now defined as “a person who hasn’t used his/her account for a year”, the model may not be able to recognize the real churners. This will increase the number of False Negative (FN) and False positive (FP) and consequently lower the level of model’s accuracy.

In the first stage of the empirical analysis, K-Means clustering technique was applied the training set in WEKA. As a result, different numbers of clusters were generated, in order to choose the best number of clusters (that is: 2 clusters, 3 clusters or 5 clusters etc.) which models the problem. It was discovered that the result which gave five clusters was better than the others. In the second stage of the analysis, the chosen clusters were then analyzed using the RIPPER (JRip) classification algorithm. The clustering algorithm was therefore used to partition the customers into groups with similar clustering characteristics based on their attributes as shown in Table 3.
Table 3: Characteristics of 5 initially extracted clusters from the customer’s database

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Instances (%)</th>
<th>Age</th>
<th>Gender</th>
<th>Balance</th>
<th>DormtStatus</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>183(37%)</td>
<td>57</td>
<td>Male</td>
<td>6265.51</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>26(5%)</td>
<td>50</td>
<td>Female</td>
<td>25783.1569</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>117(23%)</td>
<td>62</td>
<td>Male</td>
<td>18139.4779</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>76(15%)</td>
<td>32</td>
<td>Female</td>
<td>10601.4605</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>98(20%)</td>
<td>34</td>
<td>Male</td>
<td>20712.7888</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 3 shows that behavior of Female customers who were clustered in groups 1 and 3 while that of Male customer was clustered in groups 0, 2, and 4. The transactional characteristics of the customers in each of the groups are presented as follows:

a. **Cluster 0 – Cluster Zero consists of 183 cluster instances, representing 37 percent of Male bank customers in their late 50’s with operating account balance above 6000 NGN, who are at high risk of Churn based on their account dormancy status identified as being Inactive;**

b. **Cluster 1 – Cluster One consists of 26 cluster instances, representing 5 percent of Female bank customers in their early 50’s with operating account balance above 25000 NGN, who are at no risk to Churn based on their account dormancy status identified as being Active;**

c. **Cluster 2 – Cluster Two consists of 117 cluster instances, representing 23 percent of Male bank customers in their early 60’s with operating account balance above 18000 NGN, who are at no risk of Churn based on their account dormancy status identified as being Active;**

d. **Cluster 3 – Cluster Three consists of 76 cluster instances, representing 15 percent of Female bank customers in their early 30’s with operating account balance above 10000 NGN, who are at no risk of Churn based on their account dormancy status identified as being Active; and**

e. **Cluster 4 – Cluster Four consists of 98 cluster instances, representing 20 percent of Male bank customers in their early 30’s with operating account balance above 20000 NGN, who are at no risk of Churn based on their account dormancy status identified as being Active.**

One interesting way to examine the data in these clusters is to inspect the clusters visually through Visualize cluster assignments in WEKA. In order to demonstrate in a chart how the clusters are grouped in terms of Age and Dormancy Status, the chart was plotted with Age (Num) on the X axis, Dormancy Status (Num) on the Y axis, and the Color to Cluster (Nom) with the “Jitter” turned up completely to artificially scatter the Plot Points for enhanced visualization as shown in Figure 4.

**Figure 4: Cluster Visual Inspection of Age(Num) against DormtStatus(Num)**

3.3 RIPPER (JRip) Classification Analysis

In order to evaluate the developed model, RIPPER (JRip) classification rule based algorithm was used to generate sets of rules and performance evaluation metrics that define the goodness of fit of the developed model. At this stage of empirical analysis, two criteria are used based on the confusion matrix. Criteria one is based on the Accuracy rate represented by equation 5, which identifies the percentage of the total number of predictions that were correctly classified; while criteria two is based on the Actual Churners’ Rate which identifies the percentage of Churners that were correctly identified. However, to provide more robust and better indication of how well the developed classification through clustering model will perform when trained on new data, 10-fold cross validation was applied. Figure 5 shows the WEKA Knowledge Flowchart for the JRip classifier algorithm.

![Figure 5: JRip algorithm Knowledge Flowchart](image)

Using 10-folds cross-validation, the dataset is split into 10 mutually exclusive subsets, and then each subset is used as test set while the remaining datasets are used as training set. This procedure is repeated 10 times, where one fold is used for pruning, and the other folds are used in growing the rules, example of which are shown in Table 5.

<table>
<thead>
<tr>
<th>#</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(age &gt;= 62) and (balance &lt;= 1357.282557) =&gt; dormtStatus=Y (111.0/18.0)</td>
</tr>
<tr>
<td>2</td>
<td>(balance &lt;= 22923.25813) and (balance &gt;= 1169.263145) and (gender = Male) and (age &gt;= 34) and (age &lt;= 45) =&gt; dormtStatus=Y (28.0/6.0)</td>
</tr>
<tr>
<td>3</td>
<td>=&gt; dormtStatus=N (361.0/85.0)</td>
</tr>
</tbody>
</table>

4. CONCLUSION

According to [24], customer churn analysis has become a major concern in almost every industry that offers products and services. In this study a data mining model that can be used to predict which customers are most likely to churn or switch their banks was developed. The model used K-Means clustering in the first stage and a rule-based algorithm (JRip) in the second stage. The model was developed using case data from one of the major banks in Nigeria. The developed model can provide banks with useful knowledge regarding customer transactional behavior, help banks to identify likely churners and hence develop customer retention modalities. Bank customer databases contain a lot of information which if appropriately analysed can further enhance banking practices and operations. Future work will also focus on fraud detection and marketing and enhancement of banking operations and practices.
REFERENCES


Author’s Biographies

Dr. A. B. Adeyemo is a Lecturer at the Computer Science Department, University of Ibadan, Oyo State, Nigeria. He has a PhD, M.Tech and PGD in Computer Science from the Federal University of Technology Akure and B.Sc (Hons) in Engineering Physics from the University of Ife (now Obafemi Awolowo University) Ile-Ife. His research interests are in Data Mining, Computer Networking and Internet Computing. Email: sesanadeyemo@gmail.com

Oluwaseyi Oyeniyi is a student of University of Ibadan, Oyo State, Nigeria. She hold B.Sc, M.Sc Computer Science and Professional certifications and trainings CCNA, ITIL, PMP, CISA. She has banking experience with deep understanding of core banking application Oracle FCUBS v.12.0 and Oracle FCDB.