A Systematic Review of Consumer Behaviour Prediction Studies

Afolabi Ibukun.T1, Olufunke Oladipupo1, Rowland E. Worlu2 & Akinyemi I. O1.

1Department of Computer and Information sciences, Covenant University, Ota, Nigeria
2Department of Business Management, Covenant University, Ota, Nigeria

Abstract: Due to the importance of Customer behaviour prediction, it is necessary to have a systematic review of previous studies on this subject. To this effect, this paper therefore provides a systematic review of Customer behaviours prediction studies with a focus on components of customer relationship management, methods and datasets. In order to provide a comprehensive literature review and a classification scheme for articles on this subject 74 customer behaviour prediction papers in over 25 journals and several conference proceedings were considered between the periods of 1999-2014. Two hundred and thirty articles were identified and reviewed for their direct relevance to predicting customer behaviour out of which 74 were subsequently selected, reviewed and classified appropriately. The findings show that the literature on predicting customer behaviour is ongoing and is of most importance to organisation. It was observed that most studies investigated customer retention prediction and organizational dataset were mostly used for the prediction as compared to other form of dataset. Also, comparing the statistical method to data mining in predicting customer behaviour, it was discovered through this review that data mining is mostly used for prediction. On the other hand, Artificial Neural Network is the most commonly used data mining method for predicting customer behaviour. The review was able to identify the limitations of the current research on the subject matter and identify future research opportunities in customer behaviour prediction.

Keywords: Consumer Behaviour, Prediction, Statistics, Data Mining, Dataset, Customer Relationship Management, Literature Review
1. Introduction
Consumer behaviour can be defined as the study of individuals, groups or organizations in a bid to understand the process of their selecting, securing, using and disposing the products, services, experiences or ideas (Raorane & Kulkarni, 2011). Consumer behaviour in the context of this research also includes consumer’s loyalty and customer churn. Consumer loyalty can be defined, according to East et al. (2005) as repeat patronage behaviour which is the combination of attitude and behaviour. In industrial and service marketing, behavioural loyalty is viewed as retention of the brand (Reichheld 1996; Reinartz and Kumar 2000). Customer churn, also known as customer attrition or customer turnover, is the loss of existing customers to another company or service provider (Kerdprasop et al., 2013).

It is important to predict customer behaviour because; the knowledge of a customer’s loyalty would be useful for improving CRM. It will also help in customer model-building process and evaluating the results of CRM-related investments (Buckinx, 2007). Furthermore, it will help to improve the success rate of acquiring customer, increasing sales and establishing competitiveness (Qiu, 2014).

There are some literature reviews in the field of customer behaviour studies which include the following: In 2009, Ngai et al., carried out an academic literature review of the application of data mining techniques to CRM. Their study was able to classify Customer Relationship Management along the following dimensions; Customer Identification, Customer Attraction, Customer Retention and Customer Development. They also identified Association, Classification model as the most commonly used model for data mining in CRM. Brosekhan in 1995 did a review on the consumer buying behaviour to reveal two broad paradigms, the positivist and the non-positivist. The positivist paradigm encompasses the economic, behavioural, cognitive, motivational, attitudinal, and situational perspectives, while the non-positivist paradigm, envelops the interpretive and postmodern perspectives. A study to investigate the relationship between customer complaints behaviour, complaint handling mechanisms and customer loyalty was carried out by (Komunda, 2013). It was discovered that a growing body of literature suggests that customer loyalty has positive impact on client retention.

In 2003, a systematic review on online consumer behaviour was carried out by Cheung et al., and a research framework with three key building blocks (intention, adoption, and continuance) was proposed. The findings show that factors affecting intention of buying from the web is the main focus on the existing research using TRA (Theory of Reasoned Action) and its related theories as the method. (Dahiya &Talwar, 2015) performed a state-of-art review of various methods and researches involve in churn prediction and concluded that Customer churn has been identified as a major problem in Telecom industry and aggressive research has been conducted in this by applying various data mining techniques.
Looking through all the systematic reviews in literature till now, there has not been much attention paid to review of literature on predicting consumer behaviour holistically. In order to fill this gap and because of the importance of customer behaviour prediction, this paper is carrying out a systematic review on customer behaviour prediction studies with a focus on components of customer relationship management, methods and datasets.

2. Research Questions

RQ1: Which journal is dominant in the field of customer behaviour prediction?
The main motivation for this is to discover most important customer behaviour prediction journal.

RQ2: What are the datasets mostly used for customer behaviour prediction?

RQ3: What kind of datasets are the most used for customer behaviour prediction after year 2010?
The main motivation behind RQ2 and RQ3 is to discover which dataset is used to predict customer behaviour and if there is change in dataset in the most recent times.

RQ4: What kind of methods are the most used for customer behaviour prediction?

RQ5: What kind of methods are the most used for fault prediction after year 2010?
The motivation for this question is to discover trends and opportunities for prediction method focus before and after 2010.

RQ6: What kind of component of CRM is focused on when performing customer behaviour prediction?

RQ7: What kind of component of CRM is focused on when performing customer behaviour prediction after year 2010?
The motivation behind this is to discover the most predicted components of CRM so that future research can explore other components.

RQ8: What is the percentage of publications published after year 2010?

RQ9: What are the limitations of current research in Customer behaviour prediction?

3. Methodology

3.1 Inclusion Criteria

Consumer loyalty can also be defined, according to (East, 2005) as repeat patronage behaviour which is the combination of attitude and behaviour. Also, Predicting churn, i.e. if a customer is about to leave for a competitor, is an important application of analysing customer behaviour (Eichinger, 2006).

Due to this definition, this systematic review on consumer behaviour prediction includes publications on predicting loyalty, churn and consumer behaviour. Therefore, the literature search was based on the descriptor, ‘customer behaviour prediction’, ‘customer loyalty prediction’ and churn prediction. Two hundred and thirty articles were identified and reviewed for their direct relevance and seventy four were subsequently selected, reviewed and classified. Papers directly related to the search criteria was included. Position
proceedings and papers which do not include experimental results were excluded. Papers with respect to their years, datasets, customer relationship management component and method have been examined.

3.2 Classification method
The result of predicting consumer behaviour is an input to improving customer relationship management. During the review, it was discovered that the purpose of customer behaviour prediction papers falls in either one or a combination of the following dimensions of CRM; Customer Identification, Customer Attraction, Customer Retention and Customer Development. This was further validated in existing literature (Kracklauer et al., 2004), (Ngai et al., 2009). Based on this, the first classification criteria is the dimension of CRM. The second classification criterion is the dataset in which this prediction experiments were based on. Finally, the last classification criteria is the method implemented in the prediction.

3.2.1 Classification framework – CRM dimensions
The first classification framework used in this study is based on CRM dimensions. According to (Ling & Yen, 2001), the four dimensions of the CRM cycle are essential efforts to gain customer insight. Gaining customers insight on the order hand is the purpose of customer behaviour prediction. For this purpose, this review classifies papers based on their purpose, which is customer identification, attraction, retention and development.

1. Customer Identification: Customer identification also known as acquisition has to do with targeting the population who are most likely to become customers. It includes target customer analysis and customer segmentation.

2. Customer Attraction: After the market has been segmented, customer attraction then helps organizations to direct effort and resources into attracting the target customer segments.

3. Customer Retention: Customer retention can be defined as the activity that an organization undertakes in order to reduce customer defections. To be successful, customer retention starts with the first contact an organization has with a customer and continues throughout the entire lifetime of a relationship involves loyalty programs, one to one marketing and complaints management (Singh, & Khan, 2012) (Ngai et al., 2009).

4. Customer development is the consistent expansion of transaction intensity, transaction value and individual customer profitability. It involves Customer lifetime value, up/cross selling and market basket analysis (Ngai, 2009).

3.2.2 Classification framework – Implementation method
The classification framework-implementation method contains 10 categories which is described below.

Statistics: In this research, this classification group is contains techniques such as auto regression, logistic regression, linear regression, structural equation modelling and so on.

Clustering: K means clustering mostly used in this category works as follows; given a set of points in a Euclidean space and a positive integer k (the number of clusters), K means split the
points into k clusters so that the total sum of the (squared Euclidean) distances of each point to its nearest cluster center is minimized (Lloyd, 1982).

**Classification:** This includes the following: Naïve Bayes; Bayesian classifiers assign the most likely class to a given example described by its feature vector (Rish, 2001). Bayesian Network, Decision Tree (DT), Support Vector Machine (SVM) are all components of the classification category.

**Psychological prediction Models:** These classifications involve the group of prediction models which has its foundation in the psychological field of study.

**boosting algorithms:** The boosting algorithms used in this classification includes Real AdaBoost, Gentle AdaBoost and Modest AdaBoost (Shao et al, 2007).

**Particle Swarm Optimization** used in this classification is an evolutionary data mining technique (Liu and Chen, 2012).

**Neuro-Fuzzy:** This classification is a combination of neural network and fuzzy logic technique in data mining.

**Graph Mining Technique:** Graph mining is a technique used to extract characteristic patterns from a variety of graph structured data (Inokuchi, 2002).

The remaining category in the method classification includes Artificial Neural Network (ANN) and Association Rule Mining.

### 3.2.3 Classification framework – Dataset

One of the greatest threat to the validity of customer behaviour prediction studies nature of data used for the prediction.

The Classification framework using Dataset is based on the following sources of the dataset used for the experiment.

1. **Questionnaire survey**
2. **Public data repository**
3. **Organisational data**
4. **Unknown (not stated by the author)**

**Questionnaire survey:** This includes data gotten by the researchers conduction a questionnaire survey or interview of the individual or environment researched.

**Public data repository:** this includes commercially available public data, publicly downloaded data

**Organisational data:** this contains data collected from an organizational database, organisational information system, for example, their website log details etc. it also includes company transactional data, data purchased from a company.

### 3.3 Threats to Validity

The authors of this review paper are academic researchers in the field of data mining, text mining, customer relationship management and marketing. They have also published paper in the mentioned field.

There was no bias in choosing the papers reviewed. The papers were not searched based on issue-by-issue or manual reading of titles of all published papers in journals, but based on the following steps

1. **Online database search.**
2. **Initial classification by first researcher.**
3. Independent verification of classification results by second researcher;
4. Final verification of classification results by third researcher and fourth researcher.

Conference papers were not excluded because it contains experience reports. are mostly published in conference proceedings.

4.0 Results

53 journal papers out the 74 papers reviewed were journals and the rest includes conference proceedings, white paper, manuscript, advance science and technology letters. Publication years of papers are between year 1992 and 201. Fig. 4.1 is a bar chart of publication year on the x-axis and the number of papers published in that year on the y-axis for papers in review. 72% percentage of papers are journal papers and the rest fall within 28%.

![Number of Publication](#)

**Figure 4.1**

The research questions is being addressed by the following section respectively

4.1 Which journal is dominant in the field of customer behaviour prediction?
Table 4.1 Most important customer behaviour prediction journals.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Journal</th>
<th>Number</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Elsevier Journal (Expert system with applications)</td>
<td>17</td>
<td>20%</td>
</tr>
<tr>
<td>2</td>
<td>Advanced Science and Technology Letters</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>2</td>
<td>International Journal of Computer Applications</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>2</td>
<td>European Journal of Operational Research</td>
<td>2</td>
<td>3%</td>
</tr>
</tbody>
</table>

According to the table above, the most dominant journal gotten from the review literatures is the Expert Systems with Applications journal. This journal has been consistently publishing customer behaviour prediction articles as featured in the review between 2005 and 2011.

4.2 What are the datasets mostly used for customer behaviour prediction?

![Figure 4.2](image)

According to the pie chart in Figure 4.2, the Organizational database (69%) has been the main source of data for customer behaviour prediction since 1992, followed by the data gotten from the questionnaires administered to relevant respondents (20%). Public data repository (7%) is not often used, this is due to the nature of the prediction focus.
Customer behaviour prediction system is always interested in using data that is directly related to the organization in question.

4.3 What kind of datasets are the most used for customer behaviour prediction after year 2010?

![Figure 4.3](image)

Figure 4.3 also shows a consistency in the type of dataset used for customer behaviour prediction. The organisational dataset (60%) does not lose its value for prediction purposes in recent times i.e. between 2010 till date.

4.4. What kind of methods are the most used for customer behaviour prediction?

![Figure 4.4](image)

Figure 4.4 Bar chart of number of publications against method.

Form figure 4.4 above, classification technique (combination of Support Vector Machine (SVM), Naïve Bayes, Bayesian Network and Decision Tree (DT) is the prediction method with the highest count, followed by statistics in the prediction method, it should also be noted that statistics is a combination of auto regression, linear regression, structural equation modelling and logistic regression. Comparing figure 4.4.
with figure 4.5, it is observed that the trend in which the method is being adopted still remains the same even after the year 2010.

Table 4.2

<table>
<thead>
<tr>
<th>Method</th>
<th>Method(Details)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>k means clustering</td>
<td>Vijayalakshmi et al,(2013 ) Hosseini et al.,2010, Liu and Chen (2012)</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td></td>
<td>Kirui et al,(2013 ),Qiu (2014)</td>
</tr>
<tr>
<td>Method</td>
<td>Techniques</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Association Rule Mining</td>
<td>Apriori Algorithm, market base analysis etc</td>
<td></td>
</tr>
<tr>
<td>Boosting</td>
<td>boosting algorithms</td>
<td></td>
</tr>
<tr>
<td>Optimization Technique</td>
<td>Particle Swarm Optimization</td>
<td></td>
</tr>
<tr>
<td>Graph Mining Technique</td>
<td>Liu and Chen (2012)</td>
<td></td>
</tr>
<tr>
<td>Hybrid of Clustering and Fuzzy Logic</td>
<td>Neuro-Fuzzy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Abbasimehr et al, (2011)</td>
<td></td>
</tr>
</tbody>
</table>

4.5 What kind of methods are the most used for customer behaviour prediction after year 2010?

Figure 4.5 Bar chart of number of publications against method after year 2010
In order to have a detailed explanation of the references of the methods described in figure 4.4 and 4.5, Table 4.2 list the methods discovered from the review against the reference of the publications which used this method. This classification is based on their application in the reviewed publication. In table 4.2, the classification techniques can be broken down to Support Vector Machine (SVM) with 5 publications, Naïve Bayes with 3 publications, Bayesian network with 2 publications and Decision Tree (DT) with 11 publications.

4.6 What kind of component of CRM is focused on when performing customer behaviour prediction?

![Component of CRM](image)

From Figure 4.6, it is obvious that the most predicted component of customer relationship management is retention (60%). This shows that organizations spend more resources in researching to retain customers, because this contributes directly to their profit. Even after the year 2010, Figure 4.7, retention (51%) is still the most researched component of CRM. As from 2010 upwards, the identification (from 12% to 16%) and customer development (11% to 16%) component of CRM is seen to gain more focus.
4.7 What kind of component of CRM is focused on when performing customer behaviour prediction after year 2010?

![Component of CRM After 2010](image)

4.8 What is the percentage of publications published after year 2010?

In this review, the percentage of publications used which their publication year falls between the year 2010 and 2014 is 47%.

4.9 What are the limitations of current research in Customer behaviour prediction?

The limitation of the current research, which gives opportunity for future research in customer behaviour prediction, includes the following:

- There is need to continuously explore new mining methods, in order to make prediction models more precise and more informative (Yihua et al., 2014).
- Using k-means approach for predicting customer behaviour still requires an extensive evaluation in order to detect if there is over fitting by means of cross validation or bootstrap methodologies (Vijayalakshmi et al, 2013)
- There remains unexplored aspects of churn prediction which has to do with prediction when possible churners are likely to quit, also there is need to improve the prediction rates. Also in churn prediction, there is need to solve the class imbalance problem, where the class we are interested in, is the minority (Kirui et al, 2013)
- Also, there is need for further research in investigating whether web data can be implemented in customer acquisition models (D’Haen et al, 2013)
- Lastly, there is need to make the customer acquisition process more targeted by integrating more textual information in the prediction process(Thorleuchter et al, 2012).

5. Conclusion

In this study, a comprehensive systematic review and analysis of paper published in journals and conference proceedings in the area of customer behaviour prediction for 15 years were conducted. The aim was to provide a
comprehensive literature review and a classification scheme for articles on this subject base on customer relationship management, methods and datasets. Through the review the following were observed:

- Research on predicting customer behaviour is ongoing and is of most importance to organisation.
- Most studies investigated customer retention prediction and this was done mostly using organizational data as dataset for prediction.
- Also, data mining methods were used more often in predicting customer behaviour, as compare to the statistical approaches and the most commonly used data mining method is Artificial Neural Network.

The above observations imply a very rich research opportunity in the area of customer behaviour prediction and also, reflect a good relationship between the research community and industries. To this effect, this study identified limitations of current researches on the customer behaviour prediction studies and identifies future research opportunities.

References


Appendix A. Papers on Customer Behaviour Prediction in Review


imported wine in city of Hangzhou”. In International Conference on Management, Economics and Social Sciences, 591-595


