RESPONSE MODELLING WITH *K*-NN AND XNB

Dr Jay B. Simha

CTO and Head, Analytics, Abiba Systems

Advisor and Professor,

REVA Academy for Corporate Excellence (RACE)

REVA University

Bengaluru, India

jay.b.simha@abibasystems.com

Dr Shinu Abhi

Professor and Director

REVA Academy for Corporate Excellence (RACE)

REVA University

Bengaluru, India

shinuabhi@reva.edu.in

ABSTRACT

Response modelling is one of the important predictive modelling techniques used to get insights into the responses or behaviour of the events like repeat purchase by customers. In this work, application of a lazy learning method called *k*-NN and a generative model called xNB are used to evaluate the data set for repeat purchase behaviour for an e-commerce business. The RFM data view on a real-world data set is used for performance evaluation. Different values of *k* for KNN are evaluated for suitability and robustness. A tree augmented naïve Bayes and the BayesNet classifiers are also explored and compared with simple naïve Bayes as the baseline. The results of the experiment are discussed with additional work planned for the project.

Key-Words: *Response, k*-NN, *Bayesian classifier, Scalability, RFM*

1.0 INTRODUCTION

According to a study by Indian Brand Equity Foundation, the Indian e-commerce market is expected to grow at a compound annual growth rate (CAGR) of 44.7 percent and, it will touch $64 billion by 2020. Electronics is currently the largest segment in e-commerce in India with a share of 47 percent followed by apparels with 31 percent. There are 1.2 million transactions occurring in the e-commerce retailing per day (IBEF report, 2017). There are approximately more than 500 e-commerce companies in India, each vying and wooing the customers with better products and offers. In order to stay afloat in the fiercely competitive market, the e-commerce companies have to have an insight into their customer's purchase behaviour. More and more e-commerce companies are turning to big data analytics (Analytics Vidya, 2017).

It is important for e-commerce companies and vendors to identify visitors who can be converted to regular and loyal buyers and then target them to reduce promotion cost and increase the return on investments. Understanding customer characteristics and requirements can improve customer loyalty and in turn decrease operational costs (Chang & Tsai, 2011; Cheng & Chen, 2009). Using analytics, these companies are able to understand their customer base and anticipate and predict the likely buyers for their next visit. This is a response modelling problem, where the outcome is a binary 'Yes/No', kind of a response. Multiple techniques are available to implement this solution (Liu et al., 2016). In this work, two classes of techniques, namely *k*-NN and Bayes classifiers are used to explore the suitability for the set of data used in the experiments.

2.0 LITERATURE SURVEY

RFM (Recency, Frequency and Monetary) data is how recently the customer has purchased from the website how often, and how much the website has earned from customers' purchases in the past. RFM model is originally developed by Hughes (1996) is one of the common segmentation methods and calculating customer lifetime value (Hughes, 1996). RFM helps the e-commerce companies to predict the monetary value each customer brings in vis-à-vis how much the company should be willing to convince the customer to visit the site and to purchase the product. E-commerce companies combine recommendation engines with RFM analysis which will help the site to choose how much discount to be offered to a product for a customer based on the expected/predicted value of that customer to the site, and the proposed value of the product to the customer (Prashar et al., 2016).

RFM analysis is a useful method to improve customer segmentation by dividing the customers into various clusters for personalising services and also to predict who is more likely to respond to promotions. There are many studies applying RFM data including Bayesian networks, Association rules, and statistical methods like logistic regression etc. (Olson et al., 2009). Chen et al., have analysed the online retail data to understand the customer profiles for marketing activities (2012).

2.1 RFM and Repeat Buying Prediction

RFM data also helps the managers to figure out how many customers will repeat their visits and how many new visitors will visit their online stores in a given period of time (Lee, Zufryden, & Dreze, 2001). This information is very useful while formulating pricing and promotional strategies to improve the customer response rate (Birant, 2011; Chang & Tsai, 2011). The scope of this study is to predict repeat visits of customers from RFM data.

Lee et al have proposed a repeat visit model based on Negative Binomial Distribution (NBD), to predict the repeat visit behaviour of visitors to websites. Kanyakumari et al. (Venkatesh & Dash, 2012) have conducted experiments on multiple data sets for e-commerce repeat purchase behaviour and conclude that the NBD model is quite effective in predicting the repeat purchase (Lee et al., 2001).

Amine et al have developed a segmentation model based on k-means and LRFM (with the length of tenure as an additional variable) to identify potential customers. However, the approach will not identify the propensity of the customer for next visit (Amine et al., 2015).

Yao et al., (2014) have used visual segmentation using SOM and response to repeat visit modelling. The framework was used on a consumer base of one million and has drill-down capability to understand the insights for each of the derived segments.

Liu et al., have presented their case of a winning entry in a KDD competition for e-commerce response modelling. They have used feature engineering as the core and XG boost as the modelling algorithm. They provide the insight that no single feature is useful as a complete predictor. They also suggest that a large number of feature be developed for solving response modelling problem (Liu et al., 2016). Parashar et al., propose a neural network model to predict the online buying behaviour of Indian consumers and reported the factors affecting the online shopping behaviour. Authors reported an accuracy of 97%, but not mentioned the scalability of the neural network based approach (Prashar et al., 2016).

It appears from the literature survey, no baseline framework or model performance results for a scalable technique for online repeat buying behaviour prediction. Authors conclude that the NBD model will give insights into user behaviour and demographics. Hence in this work, we propose a baseline framework to compare the accuracy and the scalability. The criteria we have chosen is based on time and space complexity, instead of just performance on a given data set.

The *k*-NN and Bayesian approaches are used in this research to propose the framework for baseline performance.

2.2 Classifier methods

k-Nearest-Neighbour (*k*-NN) is one of the basic and fundamental classifier techniques and often used as the first choice classifier especially when there is little or no prior knowledge of the general distribution of the data. The *k*-Nearest-Neighbour classification was developed especially to address problems related to non-reliable or difficult to determine parametric estimates of probability densities of a distribution. *k*-NN is often called as a *non-parametric lazy learning* algorithm. Since the underlying distribution is not known, it is often considered as a non-parametric technique. It is also a lazy algorithm, means is that it does not use the training data points to do any *generalization*. In other words, there is *no explicit training phase* or it is very minimal. This means the training phase is pretty fast. However, the scoring phase is pretty slow due to the lazy evaluation approach of *k*-NN. Fix and Hodges introduced a non-parametric method for pattern classification that has since become known the *k*-Nearest Neighbour rule (Fix & Hodges, 1951; Cover & Hart, 1967; Devijver, 1979; Hart, 1968).

Naïve Bayes is another class of generative classifier algorithm, which assumes the independence of the attributes. The extended Naïve Bayes converts all continuous variables into categorical variables through selective discretization and selects the features automatically. Due to its low computation complexities, it has good scalability. The naïve Bayes classifier by default uses Gaussian distribution for continuous values (Dimitoglou, Adams, & Jim, 2012; Patil & Sherekar, 2013; Rish, 2001).

BayesNet is an extended Naïve Bayes (xNB), where the features are discretized and the probability is computed based on the new discretized dataset. Like Naïve Bayes, Bayes net has good scalability due to its single pass approach. The additional overhead of discretizing the continuous attributes can be automated or pre-processed (Muralidharan & Sugumaran, 2012).

NBTree is a standard decision tree. But instead of predicting the majority class it uses a naive Bayes classifier in the leaves to make the prediction. This results in as many naive Bayes

classifiers as the number of leaves in the generated tree. An advantage of NBTree classifier is its implicit feature selection inherited from decision trees (Kaess, Ila, Roberts, & Dellaert, 2010).

3.0 METHODOLOGY

The RFMTP data for this project comes from simulation and a publicly available data set from (Chen et al., 2012), which provides the transaction data for an online retail shop. In this study, the transaction data was transformed to an aggregate at the customer level and the k-means algorithm was used to segment the customer base. Subsequently, tree-based profiling of the segments was carried out to get insights into customer profiles. The main objective was to identify the best classifiers like *k*-NN and xNB for repeat visit prediction and also identify their performance at much larger scale.

4.0 EXPERIMENTS

In the simulation, an RFMTP set of variables with the response has been generated using the simple rules:

- Very recent customers (R < 45 days) is likely to visit
- Very frequent (F >7 days) and recent (45<R<90)

Randomness is added to make sure the exact rules/pattern will not be discovered by the learning algorithms. The original data to define the distributions are taken from standardizing the data set from (Chen et al., 2012) and resampling with replacement.

The simulated data for 10K, 100K, 1M and 5M are generated with a 10% response and used to test the scalability of the *k*-NN and Bayesian models used in this work.

The customer level RFM data is extracted for analysis and it has been observed that around 4381 customers are visible in the transaction history. Around 400 customers were purchasing online in the month of observation, making it around 10% active base ration at the point of analysis.

The recency plot (Fig.1) follows the expected reverse exponential curve and confirms, more people are active recently. The objective of the data mining is to increase the recency of the customers for purchase. It also shows the long-held insights that the recency is the better indicator of response.

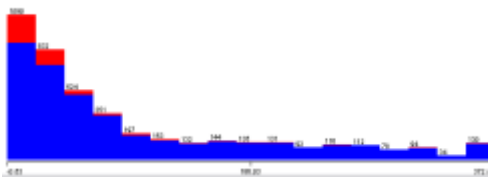

Figure 1. Recency distribution of label

Frequency also follows the expected curve as shown in Fig.2. However, since most of the customers are visiting less than average, it is necessary to differentiate different types of customers, to improve the visit frequency.

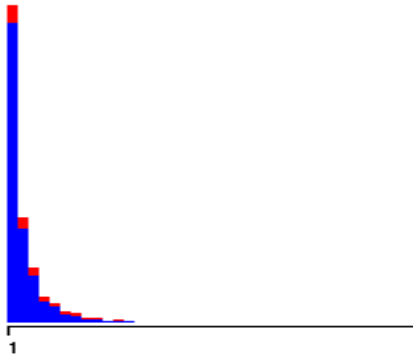

Figure 2. Frequency distribution of label

The monetary value distribution is shown below in Fig. 3. It can be seen that a high number of low-value purchases have been made and the repeat purchase ratio is very low in this segment. Hence, monetary values alone cannot be used for prediction.

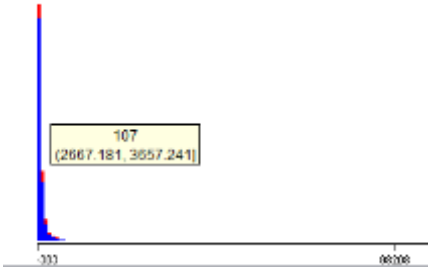

Figure 3. Monetary distribution of label

The repeat purchase odds are around 10% (Fig.4). This indicates that at the time of analysis, only 10% of the customers have done repeat purchase. The problem of this study is to identify the lookalikes of the buyers from the RFM behaviour, to promote for increasing the repeat purchase among non-visitors/purchasers.

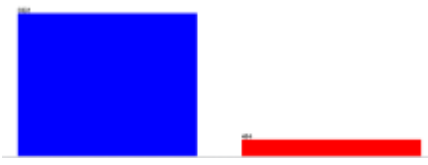

Figure 4. Class distribution of label

As suggested by Liu et al. (2016), a measure of the classification power of our approach is to compare the predicted scores from different classifiers of those customers who eventually purchase during the forecast period with the predicted scores of those who do not.

All the data is used for training, with a 10 fold cross-validation. *k*-NN (with K=1,3,5) are used to evaluate the model. Odd number values for *k* are taken to reduce the 'tie' problems likely to

occur with the even-numbered *k* values. Naïve Bayes, NBTree (Augmented Naïve Bayes) and BayesNet (xNB) models are also tested with the same methodology and compared.

5.0 RESULTS AND DISCUSSION

The values of the classification accuracy are shown below. It is surprising to see the simple solution like *k*-NN with k=1, is giving better results than higher values of *k*. This can be attributed to skewness in the data, which makes generalizations, weaker.

```
   a    b   <-- classified as
3517  374 |    a = No
 373   81 |    b = Yes
```

Figure 5. *k*-NN (*k*=1)

```
   a    b   <-- classified as
3723  168 |    a = No
 400   54 |    b = Yes
```

Figure 6. *k*-NN (K=3)

```
   a    b   <-- classified as
3795   96 |    a = No
 415   39 |    b = Yes
```

Figure 7. *k*-NN (K=5)

Table 1 shows that the "True Positives" (TP) and "False Positives" (FP) rates for the different values of *k* in the *k*-NN classifier. It can be observed that, even though the classifier with higher classification accuracy is available, it does poorly on identifying the look-alikes of the repeat buyers. The least accurate model provides the highest business benefits (at least more than 30% better than the best classifier).

Algorithm	K	TP	FP	CA
KNN	1	0.178	0.096	82.8
KNN	3	0.119	0.043	86.92
KNN	5	0.086	0.025	88.24

Table 1. *k*-NN Results

The results of the Bayesian Models are shown below (Fig. 8, 9, and 10). Again, it is surprising that, an advanced method like NBTree, which does the feature selection automatically, has performed poorer than the baseline. The only model which has performed better than the base

model is an extended naïve Bayesian classifier (xNB) called BayesNet, which has performed 50% better than the baseline model.

```
   a    b   <-- classified as
3798   93 |    a = No
 389   65 |    b = Yes
```

Figure 8. Naïve Bayes

```
   a    b   <-- classified as
3883    8 |    a = No
 432   22 |    b = Yes
```

Figure 9. xNB (NBTree)

```
   a    b   <-- classified as
3712  179 |    a = No
 357   97 |    b = Yes
```

Figure 10. BayesNet

Algorithm	TP	FP	CA
NaiveBayes	0.143	0.024	88.9
NBTree	0.048	0.02	89.87
BayesNet	0.214	0.046	87.66

Table 2. Bayesian Modelling results

The results of the experiments indicate that smaller values of '*k*' for k-NN algorithm will give a good accuracy of around 84%, with lower False Positive (FP) rate. For the same data set, the extended Bayesian algorithm has resulted in 88% accuracy and lower FP rate. Since the scalability of *k*-NN is subject to constraints. It is also observed that the FP rate of Bayesian approach is better than *k*-NN by a significant 30%.

6.0 CONCLUSION AND RECOMMENDATIONS

Bayesian approach due to its simple computational overheads has a lower complexity, and very scalable. It is observed that simple models like extended Bayes, performs surprisingly well on the state-of-art recommender systems like *k*-NN. These experiments support the insight that the simple models perform well on the scale. Hence the extended Bayesian model is used as the response modelling techniques when used at scale. These experiments support the insight that the simple models perform well on the scale.

REFERENCES

Amine, A., Bouikhalene, B., & Lbibb, R. (2015). Customer segmentation model in E-commerce using clustering techniques and LRFM model: The case of online stores in Morocco. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9*(8), 2000-2010.

Birant, D. (2011). Data mining using RFM analysis. *Knowledge-oriented applications in data mining* () InTech.

Chang, H., & Tsai, H. (2011). Group RFM analysis as a novel framework to discover better customer consumption behaviour. *Expert Systems with Applications, 38*(12), 14499-14513.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197-208.

Cheng, C., & Chen, Y. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176-4184.

Cover, T., & Hart, P. (1967). Nearest neighbour pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27.

Devijver, P. (1979). New error bounds with the nearest neighbour rule. *IEEE Transactions on Information Theory, 25*(6), 749-753.

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. *Preprint arXiv:1206.1121,*

Fix, E., Hodges, J.L. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Hart, P. (1968). The condensed nearest neighbour rule. *IEEE Transactions on Information Theory, 14*(3), 515-516.

Hughes, A. M. (1996). Boosting response with RFM. *Marketing Tools,* 4-8.

Kaess, M., Ila, V., Roberts, R., & Dellaert, F. (2010). The Bayes tree: An algorithmic foundation for probabilistic robot mapping. *Wafr,* 157-173.

Lee, S., Zufryden, F., & Dreze, X. (2001). Modeling consumer visit frequency on the internet. *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference On,* 9 pp.

Liu, G., Nguyen, T. T., Zhao, G., Zha, W., Yang, J., Cao, J., Chen, W. (2016). Repeat buyer prediction for e-commerce. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 155-164.

Muralidharan, V., & Sugumaran, V. (2012). A comparative study of naïve Bayes classifier and Bayes net classifier for fault diagnosis of the mono-block centrifugal pump using wavelet analysis. *Applied Soft Computing, 12*(8), 2023-2029.

Olson, D. L., Cao, Q., Gu, C., & Lee, D. (2009). Comparison of customer response models. *Service Business, 3*(2), 117-130.

Patil, T. R., & Sherekar, S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications, 6*(2), 256-261.

Prashar, S., Vijay, T. S., & Parsad, C. (2016). Predicting online buying behaviour among Indian shoppers using a neural network technique. *International Journal of Business and Information, 11*(2), 175.

Rish, I. (2001). An empirical study of the Naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 3*(22) 41-46.

Venkatesh, K., & Dash, M. (2012). Repeat-purchase modelling for E-commerce websites.

Yao, Z., Sarlin, P., Eklund, T., & Back, B. (2014). Combining visual customer segmentation and response modelling. *Neural Computing and Applications, 25*(1), 123-134.

WEBLIOGRAPHY

IBEF report, 2017, https://www.ibef.org/download/Ecommerce-July-2017.pdf (Last accessed on 20-10-2017)

k-NN - http://www.scholarpedia.org/article/K-nearest_neighbor (Last accessed 13-07-2017)

Analytics Vidya blogs, https://www.analyticsvidhya.com/blog/2015/08/role-analytics-e-commerce-industry, (Last accessed on 24-10-2017).

Weka - https://www.ibm.com/developerworks/library/os-weka3/index.html, last accessed 12-07-2017