Deep Temporal Features to Predict Repeat Buyers

Gaurangi Anand
Queensland University of Technology
3 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Auon Kazmi
Tata Consultancy Services Limited
2 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Pankaj Malhotra
Tata Consultancy Services Limited
14 PUBLICATIONS **111** CITATIONS
SEE PROFILE

Puneet Agarwal
Tata Consultancy Services Limited
42 PUBLICATIONS **182** CITATIONS
SEE PROFILE

Deep Temporal Features to Predict Repeat Buyers

Gaurangi Anand **Auon Haidar Kazmi** **Pankaj Malhotra**
Lovekesh Vig **Puneet Agarwal** **Gautam Shroff**
TCS Research, New Delhi, India
{*gaurangi.anand, ah.kazmi, malhotra.pankaj, lovekesh.vig, puneet.a, gautam.shroff*}*@tcs.com*

Abstract

Consumer brands often run promotional campaigns and offer discounts/coupons to attract new customers. To increase the loyal customer base and enhance return on investment (ROI), it is important for consumer brands to identify the customers who are more likely to return for repeat purchase(s) after availing an offer. The basket-level transaction history is typically used to predict the repeat-purchase behavior of customers. Different aspects of a customer's behavior can be captured using aggregate information (e.g., total number of purchases made over last one year) and temporal information (e.g., time-series of daily or weekly purchases). Most existing models either use aggregate level features only or independent window-based features without considering the sequence of transactions. We propose that a prediction model based on a combination of temporal and aggregate level models performs better compared to individual models. We use Long Short Term Memory (LSTM) as a classifier over temporal features as time-series and quantile regression (QR) as a classifier over aggregate level features. QR focuses on capturing aggregate level aspects while LSTM focuses on capturing temporal aspects of behavior for predicting repeating tendencies. The two models are then combined using a mixture of experts (ME). Experiments on a real-world Kaggle dataset demonstrate significant improvement in performance in terms of mean squared error (MSE) on using ME. Additionally, we demonstrate that the DFT coefficients of the time-series for repeaters and non-repeaters lie in separate low-dimensional embeddings indicating that there is a prominent discriminative signal in the temporal data.

1 Introduction

Often discounts, coupons, or other incentives are offered to attract new shoppers. After such promotional campaigns, it is important to identify the customers who are more likely to make a repeat purchase after the initial incentivized purchase. By focusing on these potential loyal customers in future targeted marketing campaigns, merchants can greatly reduce promotional costs and enhance the ROI. This also helps in making pertinent and useful offers to customers. Typically, the basket-level transaction history of customers is used to build models to predict the purchase or repeat behavior of customers for a product.

Many data-driven approaches have been used for analyzing and predicting customer behavior. The first step towards building models for prediction involves feature extraction from the transaction history data. Features based on recency and frequency of transactions, in addition to monetary value based features, are considered important for loyalty prediction [1, 2]. Some models extract a large number of features, and then use feature selection algorithms to obtain a subset of relevant features [3, 4]. Bayesian Neural Networks [5], Random Forests [2, 6], logit modeling [1], and Bayesian Hierarchical Model [7] have also been successfully applied for purchase behavior prediction. When data over a long time-period is available, relevant time-windows of interactions are identified and

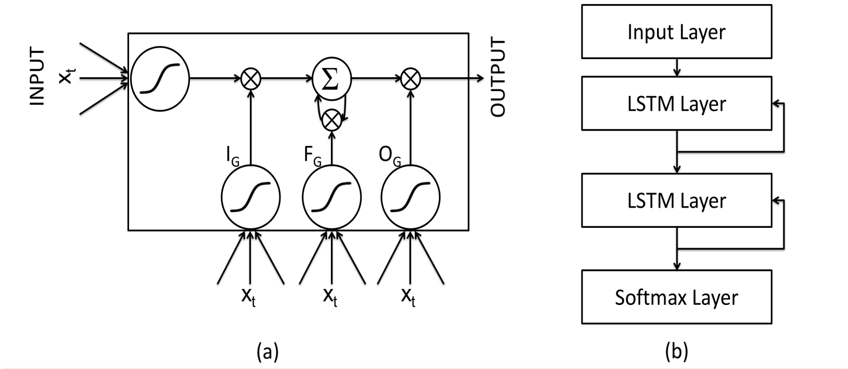

Figure 1: (a) Long Short-Term Memory Cell (b) Example of Stacked Architecture

used for learning a model based on logistic regression and bagging in combination with classification trees [8].

In order to include the temporal aspect of customer behavior, a feature's values computed over different time-windows are split into multiple features (e.g. splitting number of purchases in each of the last n months to get n different features) prior to modeling based on an ensemble of Quantile Regression, Random Forests, Gradient Boosted Regression and Neural Networks for predicting the scores [9]. In such approaches, the inherent sequential / temporal aspect of the customer behavior remains unexplored.

To the best of our knowledge, temporal models based on time-series of customer-behavior have not been used for *repeater* prediction. Repeaters are the customers who end up making a repeat purchase on the offer-product. Existing approaches use aggregate features ignoring the fact that continuous time-series data may provide additional signals that capture the behavior of the customers. Different variants of Long Short-Term Memory Networks (LSTMs) [10, 11] have been shown to give state-of-the-art results on many sequence classification tasks such as phoneme classification [12], emotion recognition [13, 14], and handwriting recognition [15]. LSTMs overcome the vanishing gradient problem in Recurrent Neural Networks by utilizing multiplicative gates to regulate information flow into and out of "memory cells". The input (I_G), forget (F_G) and output (O_G) gates in a "memory cell" allow for constant error flow s.t. the contents of the memory cell are not perturbed by irrelevant inputs and outputs. LSTMs capture long-term temporal correlations along with the original capability of neural networks to learn dependence between different features to predict the outcome. Further, stacking layers of LSTM units have been shown to learn temporal dependencies at multiple time-scales [16, 17]. A typical memory cell and an example of stacked architecture are shown in Figure 1 [17]. We use a stacked uni-directional LSTM based classifier to predict repeater vs non-repeater over multivariate time-series, and combine it with quantile regression (QR) [18] based aggregate level model by building a mixture of experts (ME).

The rest of the paper is organised as follows: In Section 2, we provide details of feature extraction from basket-level transaction data (to capture aggregate and temporal behavior) and discriminative models (LSTM and QR), and how they are combined using ME. In Section 3, we present experiments and results, and conclude in Section 4.

2 Methodology

The purchase history of a customer c can be represented as $P_c = \{(b_t, m_t) : t \in T_c\}$, where b_t is the set of products (or basket) bought by customer at time t, m_t is the market (or store-chain, online site) from where the purchase was made, T_c is the set of timestamps at which purchases are made. Further, we assume availability of products' attributes such as price, category, company, brand, etc. Given this data, we extract features based on the feature types defined in Section 2.1, as time-series as well as at aggregate level for a customer-product pair, and use them to to predict whether a customer will purchase a product again after availing an initial offer on the product.

Feature Type	Feature Name	Aggregate	Temporal
Customer-Based	TotalVisits, TotalSpend, DistinctProducts, DistinctCompanies, Loyalty,..	Y	Y
Product-Based	RepeatFractionProduct, RepeatFractionProductCompany,..	Y	N
Customer-Product Interaction-Based	AmountSpent, PurchaseFrequency, QuantityBought	Y	Y

Table 1: Sample features used for QR and LSTM models. Note that product-based features are not of temporal nature.

2.1 Feature Extraction

We extract three types of features capturing different aspects of customer behavior:
Customer-based features: These features capture a customer's overall purchasing behavior in terms of total visits made, number of distinct products / brands he purchased from, loyalty of the customer, total spend, etc.
Product-based features: Some offers have more repeaters compared to others (due to reasons such as marketing strategy, discount given, quality and popularity of product on which offer is made). We consider features such as fraction of customers who become repeaters for the offer-product, and similarly, for the offer-product's brand, company, etc. after a promotional campaign (*RepeatFractionProduct* in Table 1).
Customer-Product interaction based features: These features capture affinity of a customer to the offer-product. We consider features such as the number of visits, quantity bought, and amount spent by a customer on the offer-product, and similarly, on the offer-product's brand, company, etc.

For *aggregate level behavior*, we compute the values of above mentioned features over the entire transaction history. To capture behavior at a fine-grained level, we split aggregate level features into features computed over non-overlapping time-windows. These fine-level features are used as independent features in the aggregate level model and in the form of time-series in temporal model for deep learning. We model the problem as that of a classification problem where the actual value to be predicted by the models for a customer is set to either 1 (if customer is a repeater) or 0 (if customer is non-repeater).

2.2 Model Learning

We learn aggregate level and temporal models, and use a mixture of experts over the two models as the final prediction model:

QR based aggregate level model: We use a quantile regression (QR) [18] based aggregate level model. The loss function for QR is $q(y-p)I(y \geq p) + (1-q)(p-y)I(y < p)$, where y is the actual value (label), p (=$\mathbf{w}_q.\mathbf{x}$) is the q-quantile prediction by regression ($\mathbf{w}_q$ is the weight vector and $\mathbf{x}$ is the aggregate level feature vector for a customer) and I is the Indicator function. Positive data points (repeaters) get a weight of q and negative data points (non-repeaters) get a weight of *(1-q)* which allows for dealing with class-imbalance.

LSTM based temporal model: The n-dimensional time-series for a customer c can be represented as $S_c = \{\mathbf{s}_c^{(1)}, \mathbf{s}_c^{(2)}, ..., \mathbf{s}_c^{(T)}\}$, where each $\mathbf{s}_c^{(t)} \in \mathbf{R}^n$ for t-th time-window, T is the length of time-series. Each point in a time-series is a feature vector computed over a time-window. The network consists of n linear units in the input layer, LSTM units in hidden layer, and softmax output layer. LSTM units in a layer are fully connected through recurrent connections. For stacking LSTM layers, each unit in a lower LSTM hidden layer is fully connected via feedforward connections to each unit in the LSTM hidden layer above it.

Mixture of Experts over QR and LSTM models: Given an input vector $\mathbf{x}$, a mixture of experts (ME) [19] assigns weights to predictions of models (experts) s.t. $\mathbf{y} = \sum_{i=1}^{n} p_i(\mathbf{x}) y_i(\mathbf{x})$, where n is the number of experts, $p_i(\mathbf{x})$ is the weight learnt by ME for the i^{th} expert, $y_i(\mathbf{x})$ is the prediction score for i^{th} expert, and $\sum_{i=1}^{n} p_i(\mathbf{x}) = 1$. ME model uses the predictions given by aggregate level and temporal models, and learns a weighted sum of the predictions. We use the aggregate level feature vector as the input vector $\mathbf{x}$.
The aggregate, temporal, and ME models used give probabilities of repeating a product. These can be used to classify the customers into repeaters and non-repeaters by choosing a threshold between

0 and 1. Then, if probability of a customer repeating is below the threshold, it is classified as non-repeater, and if the probability is above the threshold the customer is classified as repeater.

3 Experiments and Results

We perform experiments on a subset of the data provided on Kaggle's "Acquire Valued Shoppers Challenge"[1]. Transaction history for customers for a period of at least 1 year prior to their offered incentive with attributes such as customer-id, store chain, department, product category, product company, product brand, date of purchase, purchase quantity, purchase amount, etc. is provided. A detailed data description can be found in [9]. Features are extracted from the transaction data as described in Section 2.1. We generated 88 aggregate features for QR and 19 temporal features for LSTM. Sample features used are shown in Table 1. We build market-wise models for nine of the markets with a total of 38k customers. There are 28.8% customers from these nine markets who are repeaters. For each market, customers are randomly divided into training, validation, and test sets, s.t., the ratio of customers in the three sets is 3:1:1.

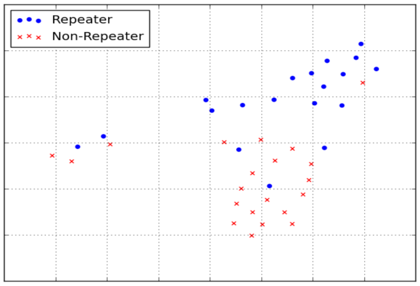

Figure 2: Sample market's t-SNE plot of DFT coefficients for cases correctly classified by LSTM and incorrectly by QR: LSTM is able to segregate cases that differ in their temporal characteristics, without relying on windowing (as a pure DFT approach would entail)

Market	**QR**	**LSTM**	**ME**	**Gain** (%)
1	0.172	0.226	0.150	12.8
2	0.197	0.220	0.180	8.6
3	0.180	0.190	0.174	3.3
4	0.363	0.259	0.248	31.7
5	0.184	0.218	0.181	1.6
6	0.172	0.217	0.165	4.2
7	0.194	0.192	0.178	8.2
8	0.294	0.256	0.216	26.5
9	0.236	0.237	0.196	20.4

Table 2: Market-wise MSE for the three models showing significant gain obtained by combining temporal and aggregate level models. Reported Gain (%) is with respect to QR.

For temporal model, each point in a time-series corresponds to weekly transactions, with resultant time-series of length 73 for each customer. We tried several deep and shallow architectures with up to 2 hidden layers (LSTM cells ranging from 5 to 25 for each hidden layer). LSTM network parameters considered are: momentum, weight decay, learning rate, and learning rate decay. Parameters considered for QR are: tau (q) and learning rate. Grid-search based parameter tuning was done on the validation set. As shown in Table 2, significant improvement can be seen in the ME model over the aggregate-level QR model. ME model has lower MSE compared to the MSEs of the individual models.

Comparison of time-series for repeaters and non-repeaters: We consider common features used in QR and LSTM models. We compute the DFT coefficients of time-series for customers that were correctly classified by LSTM and incorrectly classified by QR. The resulting 37-dimensional representation of each customer is mapped to a 2-dimensional vector using t-SNE [20]. The obtained 2-D representation of repeaters and non-repeaters for a market is shown in Figure 2. (Note: For this analysis, prediction is "repeater" if probability is above a threshold, else prediction is "non-repeater". The threshold chosen is the one with maximum F-score on the validation set.) We observe that: i) although the values for these features are very different for repeaters and non-repeaters at the aggregate level, the predictions by QR are incorrect, ii) the corresponding time-series for these samples look very different (e.g. in terms of amplitude and frequency) and are found to be useful for correct discrimination of repeaters from non-repeaters. Sample time-series for two of the features for repeaters and non-repeaters correctly classified by LSTM and incorrectly classified by QR is shown in Figure 3. Although QR can capture the amplitude aspect of the time-series at aggregate level, but it is bound to miss the other aspects of time-series (e.g. frequency).

[1] https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data

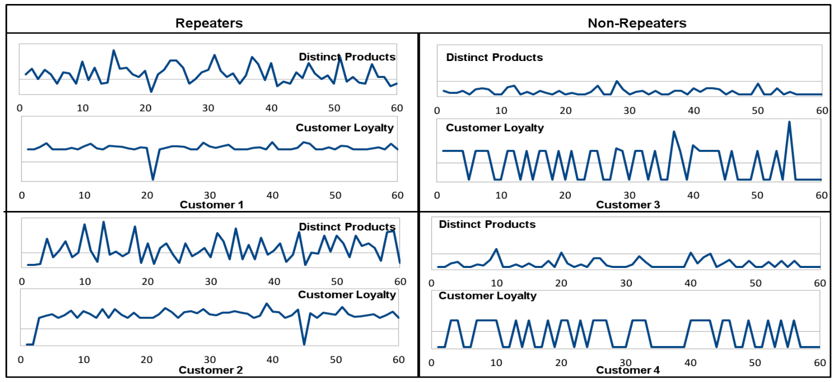

Figure 3: Sample time-series for repeaters and non-repeaters

4 Conclusion

The ME learnt over LSTM model and QR model improves over QR model in terms of MSE. Further experiments show that there are aspects of customer behavior that are missed by aggregate level features and captured by the time-series. We have studied this for purchase behavior w.r.t. products but the same approach is also applicable to other scenarios involving prediction of repeat purchases by a customer for offers on merchants, brands etc. Since LSTM training is computationally expensive, we have explored temporal features on a subset of the original data (25% of the customers). For future work we intend to examine the hidden layer activations of deep LSTMs while modeling the temporal features, and further analyzing the discriminative power of deep temporal features for multiple large datasets.

References

[1] Van den Poel, D., et al.: Predicting mail-order repeat buying: which variables matter? Tijdschrift voor economie en management **48**(3) (2003) 371–404

[2] Buckinx, W., Van den Poel, D.: Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. European Journal of Operational Research **164**(1) (2005) 252–268

[3] Van den Poel, D., Buckinx, W.: Predicting online-purchasing behaviour. European Journal of Operational Research **166**(2) (2005) 557–575

[4] Buckinx, W., Moons, E., Van den Poel, D., Wets, G.: Customer-adapted coupon targeting using feature selection. Expert Systems with Applications **26**(4) (2004) 509–518

[5] Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., Dedene, G.: Bayesian neural network learning for repeat purchase modelling in direct marketing. European Journal of Operational Research **138**(1) (2002) 191–211

[6] Larivière, B., Van den Poel, D.: Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications **29**(2) (2005) 472–484

[7] Pal, B., Sinha, R., Saha, A., Jaumann, P., Misra, S.: Customer targeting framework: Scalable repeat purchase scoring algorithm for large databases (2012)

[8] Ballings, M., Van den Poel, D.: Customer event history for churn prediction: How long is long enough? Expert Systems with Applications **39**(18) (2012) 13517–13522

[9] Nikulin, V.: On the method for data streams aggregation to predict shoppers loyalty. In: Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE (2015) 1–8

[10] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780

[11] Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm. Neural computation **12**(10) (2000) 2451–2471

[12] Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5) (2005) 602–610

[13] Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., Rigoll, G.: Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing **31**(2) (2013) 153–163

[14] Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., Narayanan, S.S.: Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In: INTERSPEECH. (2010) 2362–2365

[15] Graves, A., et al.: Supervised sequence labelling with recurrent neural networks. Volume 385. Springer (2012)

[16] Hermans, M., Schrauwen, B.: Training and analysing deep recurrent neural networks. In: Advances in Neural Information Processing Systems. (2013) 190–198

[17] Malhotra, P., Vig, L., Shroff, G., Agarwal, P.: Long short term memory networks for anomaly detection in time series. ESANN (2015) 89–94

[18] Langford, J., Oliveira, R., Zadrozny, B.: Predicting conditional quantiles via reduction to classification. arXiv preprint arXiv:1206.6860 (2012)

[19] Moerland, P.: Some methods for training mixtures of experts. Technical report, IDIAP (1997)

[20] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(2579-2605) (2008) 85