Proceedigns of the 34th European Modeling & Simulation Symposium ((EMSS)), 039
18th International Multidisciplinary Modeling & Simulation Multiconference

2724-0029 © 2020 The Authors.
doi: 10.46354/i3m.2022.emss.039

Customer Behaviour Classification Using Simulated Transactional Data

Ryan Butler[1,*], Ethan Hinton[1], Max Kirwan[1] and Abraham Salih[1]

[1]Department of Engineering Mathematics, University of Bristol, Bristol, BS8 1TR, UK

[*]Corresponding author. Email address: xy21089@bristol.ac.uk; all authors made equal contributions.

Abstract

The practice of commercial banks using transactional data to improve customer experience and profitability has become widespread, due to both the progression of a cashless society and advances in machine learning and data science, which has enabled the large-scale processing of such data. Current research has focused on reducing the risk of credit events by analysing transactional data using various machine learning techniques and, more recently, neural networks. However, the use of these techniques for the marketing of retail bank products is limited in the current literature. This paper introduces a novel technique to combine data of both *where* and *when* customers spend money, utilising a Convolutional Neural Network (CNN) classifier, trained on a visual representation of a customer's transaction history. From this, customer spending behavior in each transaction category can be labelled. Distinct clusters of accounts were also identified based on their spending habits and personality type, using two simulated transactional datasets produced via agent-based modelling and provided to us by Lloyds Banking Group, a major UK retail bank. All data supplied by Lloyds Banking Group was computer generated, synthetic data. At no time was real transactional data shared. The techniques introduced in this paper could be used by commercial banks to improve marketing strategies.

Keywords: Simulated Transactional Data; Convolutional Neural Network; Behavioural classification

1. Introduction

The use of customer data to target advertising, improve customer experience, and increase profits has become widespread across multiple industries in recent times. Companies that can gain the most from the rise of data analytics, data science, and big data are companies that inherently collect a vast amount of customer data. Retail banks sit firmly in this category; in the UK, as of 2018, there are roughly 70 million open current accounts (Treanor (2018)) and, as of 2016, there are 39.2 million card transactions per day (Best (2022)).

The transactional data collected by retail banks contains transaction dates and times, amounts, the location of transactions, the recipient, and other relevant data; and banks also have access to personal data about account holders such as address, age, gender, etc. The volume of transactions and variety of information contained in each transaction gives retail banks the resources to leverage big data processing techniques to make significant improvements in customer experience and increase profits simultaneously.

The main way in which retail banks make money is through providing loans and earning interest on them. To be able to loan large sums of money to people and businesses, retail banks require capital in the form of customer deposits in bank accounts. It is therefore logical to deduce that the more customers a bank has, and the more money customers deposit in bank accounts, the more money the bank can loan out to generate profit from interest payments.

Some retail banks also offer additional services to customers, such as credit cards and premium current ac-

counts. These products provide the bank with another source of interest payments - in the case of credit cards - and regular monthly or annual fees, in the case of premium current accounts and some rewards credit cards. It is important, however, for the bank to determine the suitability of a customer before they are offered credit or a loan, as there is the potential for a customer to default on their debts if they borrow more money than they can afford to repay. Banks usually assess customer suitability for loans and credit by performing a credit check on the customer.

There is significant potential for transactional data to be used to gain insights into customer behaviour. These insights would allow a bank to observe the general trends of their customers; the findings of which can be extrapolated to the wider population of a specific area of the country and used to direct marketing personalised to that area to maximise customer acquisition potential. Insights could also be used to group existing customers based on their spending habits and provide targeted marketing for products that have potential to be attractive to a specific group, or create the products if they do not already exist. Transactional data could also be used to gain insight into the spending psychology of a customer (e.g. whether they spend impulsively or are reserved and careful with their spending), which has the potential to be used to direct marketing of products such as credit cards and loans away from customers that are likely to have a credit application declined and towards customers that these products are suited to.

This paper will explore how projection and clustering algorithms can be used to group customers based on where they spend their money, and how Convolutional Neural Networks (CNNs) and data filtering techniques can be used to build a spending profile for customers. The effectiveness of these techniques at extracting useful insights from computer-generated transactional data, provided by Lloyds Banking Group, will also be assessed to determine their usefulness to a retail bank.

2. Literature Review

The majority of work in the field of agent-based financial modelling is focused on the simulation of financial markets and trading strategies (Poggio et al. (2001)) (Wang et al. (2018)) (Rekik et al. (2014)). There are also some works such as Lopez-Rojas et al. (Lopez-Rojas and Axelsson (2012)) and Koehler et al. (Koehler et al. (2005)) which focus on using agent-based modelling to generate transactional datasets to enhance fraud classification algorithms. However, currently there is a lack of literature focusing on using agent-based models to generate transactional datasets for the purpose of analysing customer specific transactional data.

Moreover, early work in the field of analysing transactional data for banking focused on the application of machine learning models to credit risk assessment, attempting to outperform traditional expert systems (Carter and Catlett (1987)). Further research as technology has developed has utilised neural network models and advanced hybrid models to improve performance. Khandani, Kim and Lo produced a paper in 2010 that categorised consumer bank transactions by the nature of the transaction such as: restaurant expenditure, bar expenditure, clothing expenses, etc (Khandani et al. (2010)). This data was then combined with credit bureau and account-balance data to predict credit risk. The researchers observed varying accuracy in the assigned categories but an overall improvement in accuracy of predicting credit events when using the transaction data.

Unlike these studies, however, an aim of this project is to identify clusters within the data where there are no labels present. Transactional clustering requires the clustering of categorical attributes and clustering of variable-length sets. Hierarchical methods have been a prevalent technique in the literature but scales poorly due to its quadratic complexity (Giannotti et al. (2002)). K-means clustering is a popular alternative and scales linearly with the size of the dataset. Density based algorithms assign clusters based on density neighbourhoods of points within the spatial dataspace (Zakrzewska and Murlewski (2005)). A comparison of these algorithms for bank customer segmentation found that the density based DBSCAN was an effective approach provided its parameters are tuned correctly (Zakrzewska and Murlewski (2005)).

A key consideration for clustering is determining whether the clusters should be "crisp" or "fuzzy", where the former means points can only belong to one cluster and the latter allows points to belong to multiple clusters. A related study by Holm (2018) attempted to cluster individuals using categorised transactions. The algorithms deployed for this used the crisp approach due to the inherent complexity of fuzzy clustering. Holm concludes that clustering based on categories is viable but sufficient data is required to enable accurate categories of transactions (Holm (2018)).

Another core aspect of this project is the utilisation of a neural network to categorise account time series data. The earliest neural networks were first developed by Warren McCulloch and Walter Pitts in 1943 and used "threshold logic" to imitate the process of thinking (McCulloch and Pitts (1943)). Deep learning, more recently developed by LeCun, Bengio and Hinton (LeCun et al. (2015)), makes use of multiple processing layers to better learn representations of data. Mirashk et al. (2019) use a Recurrent Neural Network (RNN) to predict customer behaviour from transactional data; customers were assigned labels for their transaction volume and the model was trained on these labels to predict transaction volume in unseen data. The RNN provides the benefit of maintaining Long Short-Term Memory (LSTM), first developed by Hochreiter and Schmidhuber (1997), which is advantageous for remembering long sequences of information (Mirashk et al. (2019)).

Convolutional Neural Networks (CNNs) are used extensively in computer vision, utilising convolutional layers to recognise objects or patterns within images. The model has recently started seeing increased use in other domains (Albawi et al. (2017)). Lv et al. (2019) apply a CNN-based detection model to transactional data in a 2019 study to learn latent trading patterns that correspond to illegal activities. In the study, a two-route CNN model is used to extract characteristics from the time series data, with the model performing better than traditional statistical models such as the IForest method (Lv et al. (2019)). The use of CNNs in bank transaction data is limited to this date in the literature, and another aim of this project will be to test its application in this field through the identification of patterns in visual representations of customer data.

The clustering of unlabelled simulated customer data and the use of CNNs to analyse this type of data in particular, both represent gaps in the literature that shall be explored in this paper.

3. Methodology

Our methodology comprises of two separate approaches. The first approach attempts to cluster customers based on *where* they spend their money, and the second approach groups customers based on *when* they spend their money, using heat-maps as a visual representation of a customer's data. These approaches were then combined to obtain a categorisation of customers based on both features. two simulated transactional datasete were utilised for this methodology (see Section 5).

3.1. Where customers spend their money

The first approach was to cluster customers based on where they spend their money. This was achieved by focusing solely on the recipient accounts for each transaction. These were categorised as either personal or commercial transactions, which relate to transactions going directly to other personal accounts, or transactions to commercial accounts, such as spending in a shop. The subset of all personal transactions was removed, as this does not provide any useful information pertaining to a customer's behaviour or personality type.

Within the commercial transactions, there existed numerous different vendors where customers spent their money. These were grouped into categories, with each category relating to the type of transaction. For instance, categories included spending on groceries and on healthcare. The total spend in each category was then calculated for each customer. These totals were then normalised to give the proportion of money spent in each category. Next, the difference between these proportions and the average of all customer's proportions was calculated, to show how much a customer spent in each category above or below the average. As well as representing accounts comparatively to each other, this method reduced the effects of having disproportionately sized categories.

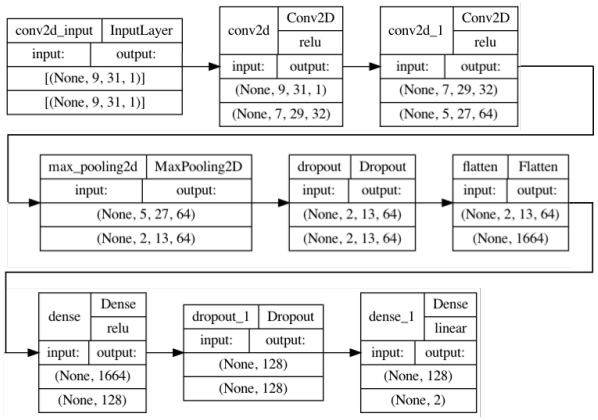

Figure 1. CNN model structure adapted from Gandhi (2018)

The next task involved grouping customers with similar spending habits together. This was achieved using UMAP (Uniform Manifold Approximation & Projection) (McInnes et al. (2018)), a dimensionality reduction technique designed for global structure preservation and high visualisation quality. UMAP uses a non-linear mapping, which makes it more applicable to this task than other, more traditional techniques such as principal component analysis. The similar data projection technique t-SNE (van der Maaten and Hinton (2008)) was also investigated, however UMAP was found to provide better clustering results, as well as being computationally much more efficient. UMAP was used to project the data down onto two dimensions, whilst preserving the relationships between the features. These original features were those described above - the difference between an account's proportional spending in each category and the average proportional spending for that category. Once reduced to two dimensions, the density-based clustering method HDBSCAN (Campello et al. (2013)) was the used to identify clusters in the data.

3.2. When customers spend their money

The transactional data for each account in Dataset 1 was converted into normalised 9×31 heat map images where each pixel is the sum of transactions for a particular day. These images were then run through the UMAP algorithm (McInnes et al. (2018)) to cluster by like characteristics in order to explore common time series spending habits. In the resulting projection, each UMAP cluster was then labelled based on the dominant features present in their constituent heat-maps.

The heat-map images were then labelled based on the dominant feature of the cluster they resided in. This was then verified manually by checking the label on each heat-map and correcting if the label was erroneous. The dataset of labelled heat-maps was then split into a training and validation set, which was used to train a CNN (LeCun and Bengio (1995)). The structure of the neural network can

Table 1. CNN model features.

Hyperparameter	Selection
Loss	Sparse categorical cross entropy
Optimiser	ADAM algorithm
Optimisation metric	Accuracy

be seen in Figure 1 and its additional parameters can be seen in Table 1.

The optimum number of epochs for this model was then determined by working out at which point the model's training accuracy becomes greater than the validation accuracy (which indicates the model is over-fitted) and choosing an epoch number slightly less than that. The model was then tested via a 10 k-fold cross validation, which is the standard (Brownlee (2020)), to determine its efficacy. Lastly, the model was then compared to a baseline classifier, which looked for trends by comparing individual pixel values.

3.3. Combining the two approaches

Following from Section 3.1, per category heat maps were generated from each account and were labelled by the neural network. Furthermore, they were also labelled according to the density of payments in that category i.e. if greater than 60% of the heat-map pixels were non-zero (indicating that the customer made payments on greater than 60% of days), the heat map was given a dense label. Otherwise, it was given a sparse label. Lastly, the per category heat-maps were further labelled by whether the non-zero heat map pixels followed a positive skew, negative skew or roughly approximated a normal distribution. This was determined by using SciPy's skew function which calculates the Fisher-Pearson coefficient of skewness.

As Dataset 2 contained account inflows and outflows to/from companies, it could be used to deduce the salary of some of the accounts. This dataset could also be broken down into per category spends. Thus, the proportion of their salary that each customer spends in each category over the entire time-period could be calculated. These values could then be used to create a population distribution of salary proportional spends for each category. Where an individual lies on this distribution, for their spending in a given category, can then be labeled depending on e.g. whether it is greater than 1 standard deviation from the mean for that category's distribution. In this case, this would be labelled as a high spender in terms of the proportion of their salary spent in said category.

Lastly, a combination of the per category heat-map labels and the salary proportion spent per category labels were used to determine the behavior of an accounts spending in each of the defined categories.

Accounts that had a salary proportion spent, for a given category, greater than 1 standard deviation away from the mean of all customers, and a negative skew as well as dense label on their individual time series heat-map, were given a impulsive high spending label for that category.

High spending labels, for a given category, were given to those accounts whose proportion of their salary spent were greater than 1 standard deviation away from the mean, and a positive/no skew to their time series data.

Average spending labels, for a given category, were given to accounts with a proportional spend within 1 standard deviation of the mean. Furthermore, below average spending labels, for a given category, were given to accounts which had a proportional spend more than 1 standard deviation below the mean.

The high, average, and below average labels were further split into dense/sparse versions depending on whether the heat-maps, for a given category, were labeled as dense or sparse. So for example, a high spender in a given category could either be classified as a high dense spender or a high sparse spender.

Finally, consecutive spending labels, for a given category, were given to accounts that possessed consecutive monthly transactions in their individual time series data. These labels were further split into consecutive high/average/below average spending depending on the proportion of the individuals salary spent in that category.

4. Data Simulation

The data used within this project was supplied by Lloyds Banking Group. Two datasets were provided, generated by Lloyds Banking Group's proprietary in-house agent-based model (ABM) of retail-customer financial activity.

The simulation for Dataset 1 was based on a town with 14,301 unique agents, which represents a proportion of the town similar to Lloyds' market share within the average UK town. The town was modelled as a graph where nodes represent agents and edges represent the distance between individuals.

Dataset 2 was simulated in a similar way, with an increased number of features added to the agents and more complex behaviors implemented by them. This was also processed on a more granular timescale (minutes as opposed to days). Due to this greater complexity, a smaller number of agents (1304) were used in order to maintain a reasonable run time.

The probabilities of events and transactions in the simulation were determined by the traits of agents. These traits were randomly assigned based on parameters, which influence the desire of agents to perform certain actions, specified in the configuration of the simulation. During the progression of the simulation, the traits of agents could change to reflect natural behavioural changes of people in the real world. Whether an action is performed by an agent is determined by random chance, with each agent having a probability threshold, based on their traits, to perform a specified action.

The underlying parameters were based upon extensive knowledge of transactional behaviour from the retail banking community combined with statistics from UK open data sources. No real transactional data was used to inform

the model due to ethical and data privacy concerns. Because the parameters of the model are based on extensive knowledge of customer behaviour, the transactional data generated by the simulation can reasonably be expected to approximate real world data.

5. Data Preparation

Dataset 1 comprised of several million simulated customer transactions, including a limited number of features such as the date and recipient account. Dataset 2 was later released providing further simulated transactions, with extra features such as the timestamp and account balance.

There were several data preparation steps required to complete the methodology. These are described below.

1. *Merchant Categorisation*: Commercial entities were grouped using a combination of regular expressions (e.g. entities including 'coffee' in their name would fall into the 'Cafe' category) and manual sorting. More complicated machine learning techniques could be utilised for larger transactional datasets, however this technique was sufficient for the data provided.
2. *Cluster Exploration*: Following the clustering of customer accounts, it was necessary to investigate the composition of the clusters. This was done by selecting a random sample of accounts from each cluster and visualising how much the accounts had spent in each category. Further metrics were calculated, such as the average spending in each category for each cluster.
3. *Employer/Salary Calculation*: Dataset 2 contained information regarding account inflows and outflows. From this, it was possible to calculate a customer's employer and salary. This was done by grouping and pivoting the data to form a dataframe with rows relating to each account and columns relating to the total income from each company. The employer and salary were then easily extracted, since each account only had inflows from one company.
4. *Heat-map Production*: The heat maps were produced by creating a pivot table depicting the total transactions per day for each of the 9 months (which was the total number of months in Dataset 1) for a given account and normalising; thereby producing a 9×31 array for that account. In the case of the per category heat maps, the total transactions per day were limited to only transactions which were in a given category.

6. Results

6.1. Commercial spending account clustering

Application of the method described in Section 3.1 resulted in 13 unique clusters identified in Dataset 1. These clusters are visualised in Figure 2. This number increased to 14 for Dataset 2, partially down to the fact that this dataset contained more merchant accounts, leading to a new categorisation including several new categories, such as education.

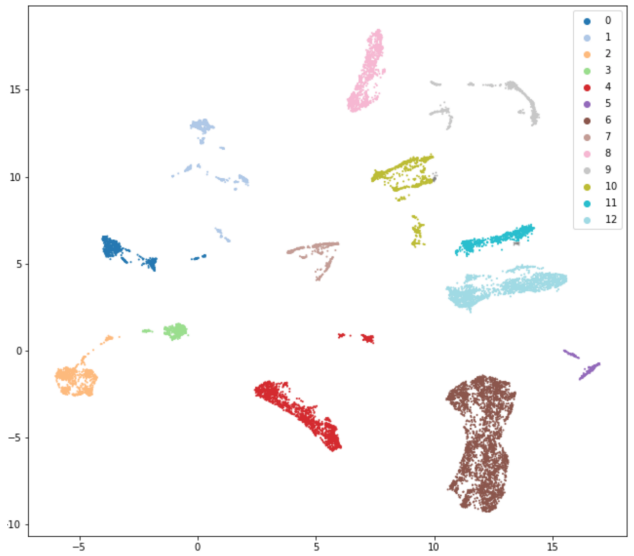

Figure 2. UMAP projection of commercial transactions from Dataset 1, clustered using HDBSCAN. UMAP hyperparameters: n_neighbors=15, min_dist=0.1, n_components=2, metric=Euclidean. HDBSCAN hyperparameters: min_samples=1, min_cluster_size=250.

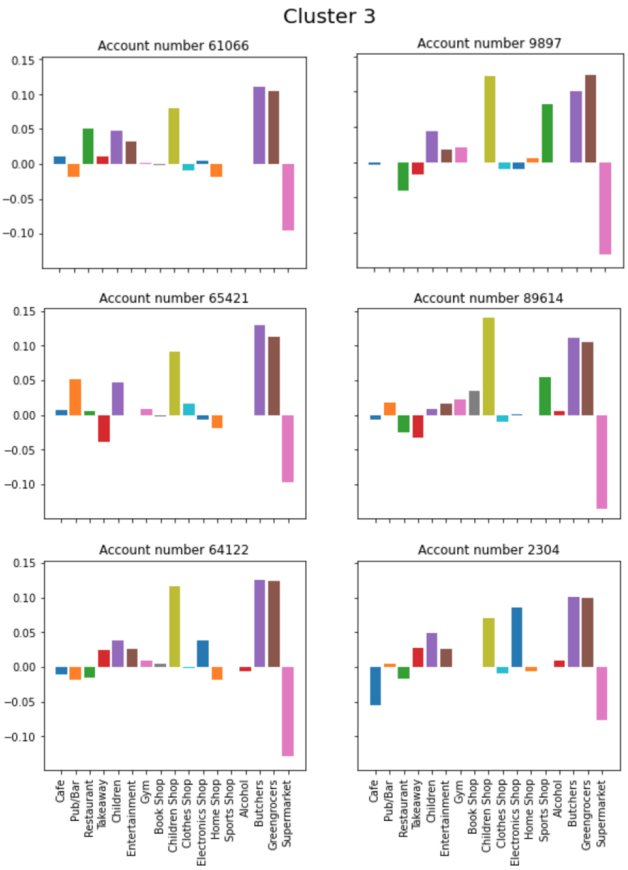

Figure 3. Representation of Cluster 3 from Dataset 1. The levels of spending from 6 random accounts have been displayed.

The clusters proved to be uniquely defined, with most clusters displaying strong trends. For example, Cluster

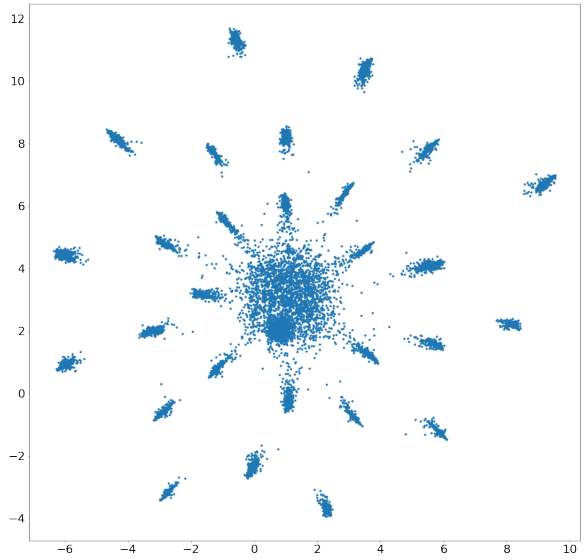

Figure 4. UMAP projection of heat-maps in Dataset 1. Hyperparameters: n_neighbors=100, min_dist=0.1, n_components=2, metric=Euclidean

3 from Dataset 1 (see Figure 3) showed a much greater than average spending within the category *children's shop*, as well as high spending at the *butchers* and *greengrocers* but low spending at the *supermarket*. This shows that customers in this cluster clearly have children and prefer to support local grocery stores rather than the big chain supermarkets, perhaps hinting at their personality type.

Trends from some of the clusters identified in Dataset 2 are described below.

- Customers whose proportional spending is well above average in both the *exercise* and *sports shop* categories.
- Customers whose proportional spending is far above average for *online shopping* and *food takeaway* services.
- Customers who spend well above average on *groceries*, and below average on *clothes* and *other shopping*.
- Customers who spend far above average on *clothes shopping*.
- Customers who spend above average on *home* and *clothes shopping* and *personal care*.

6.2. UMAP of time series spending data

The UMAP data projection in Figure 4 shows that Dataset 1's time series heat-map data can be divided into distinct clusters. From analysing each cluster's constituent heat-maps in the UMAP projection, the heat-maps can be broadly separated into two categories: spotty heat-maps (Figure 5A) and heat-maps with a line in them (Figure 5B), which indicates consecutive monthly spending.

6.3. CNN to determine consecutive monthly spending

The optimum number of epochs for the neural network defined in Section 3.2 can be determined by the graph in

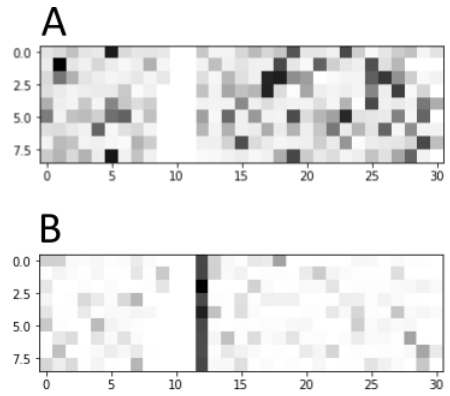

Figure 5. A: Example spotty heat-map. B: Example line heat-map.

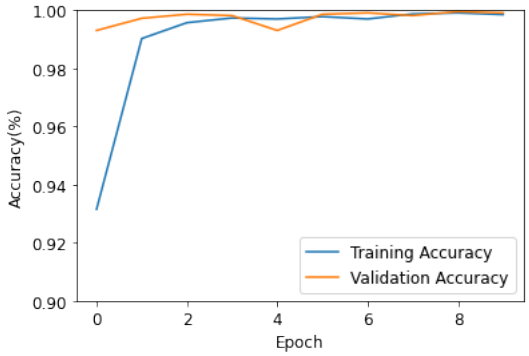

Figure 6. Epoch versus accuracy graph for CNN architecture.

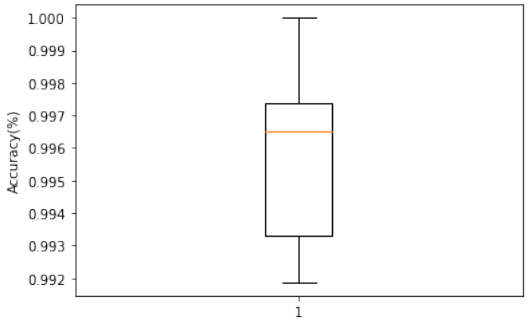

Figure 7. Boxplot showing accuracy distribution in 10-fold cross validation, with mean = 99.582 and std = 0.256.

Figure 6. This figure shows the optimum value is 2, as after this value the training accuracy exceeds the validation accuracy, implying that the model is over-fitted. Furthermore, performing a 10-fold cross validation on this CNN architecture with an epoch count of 2 yields a mean accuracy of 99.6% with an extremely low standard deviation of 0.256% (Figure 7). This performance far exceeded the baseline classifiers performance of 53.4%. Hence, this model is highly accurate at determining if consecutive monthly payments are present in a given heat map.

Table 2. Example output of the first 3 columns and 10 rows of the per category labels from the spending type classification algorithm.

Account num	Finances	Entertainment	Personal Care
999118642	C&A	0	0
998523058	C&H	H-S	A-D
996805831	C&A	C&A	A-D
995678156	C&A	H-S	0
992381583	C&A	0	0
991509944	C&A	C&A	0
991462486	0	A-S	I
990438731	C&A	A-S	0
988743819	C&H	C&A	0
986910751	0	H-D	A-D
986577456	C&A	C&A	0

Table 3. Per category spending labels and their meanings.

Label	Meaning
I	Impulsive high spender
H-D	High dense spender
H-S	High sparse spender
A-D	Average dense spender
A-S	Average sparse spender
B-D	Below average dense spender
B-S	Below average sparse spender
C&H	Consecutive monthly high spender
C&A	Consecutive monthly average spender
C&B	Consecutive monthly below average spender
0	No spending

6.4. Spending Type Classification

Table 2 and Table 3 show an example output and corresponding label definitions, respectively. The explanations of each of these labels can be found in Section 3.3. These figures show that for a given accounts transactional history, an in-depth analysis of their spending per category can be determined using a combination of time series data and total proportional spending relative to their salary, which can be used to deduce overall customer behaviour.

7. Insights

The results in Section 6 show that the methods discussed in Section 3 can be used to group similar customers together based on where they spend their money and create a spending profile for individual customers based on when and how much money they spend.

There are many potential uses for such methods from the perspective of a retail bank, however this section will focus on analysing their usefulness in directing targeted advertising of existing products to particular customers, creating new products that will appeal to large groups of people, and identifying customers that could be at a greater risk of financial trouble based on their spending habits. The ethical appropriateness of these methods will also be discussed.

7.1. Targeted advertising and product creation

Whilst retail banks are limited in the amount they can diversify the core function of their products from competitors, as the vast majority of banks offer current accounts, credit cards, personal loans, etc., they can bundle additional perks and incentives with their products to distinguish them from offerings from competing banks. Examples of incentives offered by Lloyds Bank, a UK retail bank (Lloyds Banking Group), in their Club Lloyds current account are cinema tickets, a gourmet food membership, a free magazine subscription and free movie rentals. The objective of these incentives is to attract new customers to the bank and to up-sell existing customers onto products that are more profitable for the bank.

The results shown in Section 6.1 show that, using the UMAP clustering method described in Section 3.1, customers can be grouped effectively by where they spend their money. This information would allow a bank to target advertising for a premium current account, for example, to existing customers that have a, less profitable, standard current account and already spend money on the perks offered as part of the premium offering. If targeted advertising to existing customers was deployed in this way, it could increase the adoption rates of premium current accounts whilst improving the customer experience for all customers by only showing advertisements that they are interested in.

Analysis of the spending habits of each output cluster could also help banks to develop products that appeal to audiences currently not catered to by current product offerings. For example, Figure 3 shows accounts in a specific cluster that have higher spending in the *children* category and a tendency to spend more in *greengrocers* and *butchers* rather than *supermarkets*. If a significant amount of customers fall into this cluster, it would be worthwhile for the bank to offer a premium account, credit card, or other product that offers benefits related to children (e.g., clothing discounts, toy shop discounts, etc.) to cater to the needs of this group of people and increase the number of customers using more profitable products.

The ability to profile a customer based on the frequency and amount of spending in a certain category, using the method described in Section 3.3, could also be useful to a retail bank for targeted advertising and product creation purposes. This is because if a customer consistently spends money in one category (e.g. every week or every month), it is more likely that a product offering monthly benefits (e.g. free cinema tickets every month) would be attractive to them, whereas a customer that has inconsistent, high spending in a category, might be more attracted to a product that offers one-off perks that provide significant value.

7.2. Creating a customer spending profile

Results from Sections 6.3 and 6.4 show that it is possible, using the methods described in Sections 3.2 and 3.3, to

construct a spending profile of each individual customer based on their spending habits over time. This could allow a retail bank to determine whether a customer spends or saves a large proportion of their income, spends frequently or infrequently, spends impulsively, etc.

One instance in which this could be useful to a bank is black-listing certain customers from receiving advertisements or offers for products that are financially unsuitable. An example of this would be advertising a rewards credit card that has an annual fee of £500 to a customer that has an annual salary of £20,000 as it is extremely unlikely that they will benefit from such a product. Another example would be advertising a credit card with a high credit limit to a customer that has impulsive spending tendencies and spends the majority of their annual salary, rather than saving; this would not be suitable as there is a reasonable chance that the customer would risk getting into financial trouble when using such a product and would likely not pass the required credit check on application anyway. Using the customer profile to first filter which products can be advertised to certain customers would ensure that customers are only advertised suitable products and could reduce instances of customers failing credit checks for unsuitable loans and credit cards which, in turn, could improve customer satisfaction.

As well as black-listing customers that are unsuitable for certain products on a financial basis, the customer spending profile could also allow the bank to target advertising of products that are most profitable (annual fee credit cards and loans) to customers that have high salaries and consistent high spending. The more of these types of customers that use these products, the more profit the bank will make.

7.3. Ethical considerations

Whilst a potentially useful tool for a bank to target advertising and acquire new customers, compiling transactional data to profile a customer and infer information such as annual salary and personality type could be an ethical concern. To open a current or credit card account, customers must provide personal information such as their name, age, and address. It is therefore logical for a customer to expect a bank to be able to use the personal information that they have willingly provided for other means; it is not, however, immediately clear that the bank can use this information in combination with transactional data to obtain a detailed profile about the customer's personality and tendencies. Because of this, it could be more ethically acceptable to present the customer spending profile feature as an opt-in or opt-out service that benefits both the retail bank and customer, or make it clear how transactional data will be used by the bank.

8. Future Work

The insights gained from the previous section and the methods described in Section 3 were effective in achieving the stated goals of this study. However, some considerations are necessary for the improvement of the methods discussed.

UMAP provided an effective means of grouping together customers based on the features in the dataset. The drawback of the system is that UMAP is a discriminative model and hence requires retraining on the entire dataset in order to include an additional data point within the clusters. A generative model would address this issue by learning the underlying distributions and how the data was generated. This would enable the model to scale more easily, as retraining and classifying new data points would become less computationally expensive.

The second dataset provided by Lloyds contained additional features such as customer bank balances that enabled a more sophisticated model to be developed. Other studies have used datasets that contain other features such as the cities where transactions occurred (Lv et al. (2019)), age, credit score, as well as more information about customer traits. Such features could significantly improve the classification of customer behaviour and allow for a greater degree of separation between clusters that are generated.

The CNN described in this paper was trained on a 9×31 heat-map representing the 9 months of daily transaction data. A natural extension for the model would be the utilisation of a 52×7 heat-map to see weekday variation in customer spending. A feature that could further extend the CNNs capabilities in the identification of visual patterns is the inclusion of time-stamps for each transaction. This could enable the creation of multi-dimensional heat-maps that capture total transactions by month, day, and time. This representation of spending could then be used to train one or perhaps multiple CNNs as in (Lv et al. (2019)); thereby allowing for the identification of more complex customer spending patterns.

The applications of such a model are varied, but a possible use case is to see patterns of impulsive spending at a more detailed level i.e. if a particular time in the day influences spending behavior. This could provide greater insights on what advertising would be suitable for a given customer, as well as aiding in the identification of possibly fraudulent activity where outliers appear in spending.

Another avenue for further work would be the utilisation of a larger and more diverse dataset where the user of the model could implement the model at greater scale. Furthermore, increasing the diversity and size of the training data would improve the robustness of the model at handling real-world transactional data and identifying clusters of customers with similar spending habits.

One issue with the standardisation process discussed Section 3.3 is that the amount spent in certain specific categories might not rise linearly with income. Take the example of supermarket expenses; it is not logical to expect that a person with a higher income would purchase more

food than someone on a lower income just because they have a higher income (although some correlation could be present due to the ability to purchase more expensive food products). Because of this, it could be expected that as income increases, the value for the proportion of spend in this category would decrease, as the absolute expenditure in that category would plateau above a certain income. This phenomenon is unaccounted for in the standardisation process in this report. An improvement to the standardisation process could be to separate accounts into a fixed number of bins based on income; then the proportional spending could be standardised, in each category, based only on the mean and standard deviation of spending proportions in that specific bin. This would reduce the effect of this phenomenon on the standardisation process and allow customers of varying incomes to be compared more accurately on their behaviour.

9. Conclusion

The two stated goals of this study were to use simulated data to design a model that could cluster customers based on *where* they spend their money and *when* they spend it. Through the combination of clustering based on customer spending across different categories and the training of a CNN on heat-maps representing customer spending habits across time; this study was successful in accomplishing these goals.

From the results discussed in Section 6, customers were able to be grouped into well-defined clusters based on where they spent their money, with unique trends identified in each of the clusters obtained from the dataset. The CNN achieved a 99.6% mean accuracy in identifying spending patterns across time, from the heat-maps generated for each customer. These results strongly support the novel technique utilised in this paper for the identification of trends and classification of customers through visual representations of transactional data, rather than the traditional approach of using numeric data to obtain such results. The possible extensions to the utilisation of CNNs in this field, discussed in Section 8, provide grounds for further research and implementation of the methods applied in this study.

In a commercial context, the findings of the paper are promising for its potential applications in targeted advertising and product creation that cater to specific segments of the market. Moreover, from Section 7, the identification of clusters based on spending habits provided insights on the impulsivity of customers that can be used in conjunction with known information about customer income by a bank to determine suitable products, services, or advice that are relevant for such individuals.

The novel contributions of this paper, from the EMSS perspective focused on modelling and simulation, are twofold: first, we have demonstrated that commercially valuable research advances can be made by applying contemporary machine learning techniques to synthetic data, i.e. data produced by a trusted agent-based model; and second we have shown that this combination of techniques can automatically extract commercially valuable models of underlying customer-types from the synthetic data—that is, our system *automates* the modelling process. Therefore, this novel process offers a wealth of opportunities for further research in the M&S field through the automatic modelling of unique customer behaviors.

Acknowledgements

Thanks to Marie Anderson and her colleagues at Lloyds Banking Group and to Zahraa Abdallah at the University of Bristol for their help and guidance in the early stages of this project. Thanks to Dave Cliff at the University of Bristol for comments on an earlier version of this paper.

References

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6.

Best, R. d. (2022). Card payments per day uk. https://www.statista.com/statistics/719708/card-payments-per-day-forecast-united-kingdom/.

Brownlee, J. (2020). A gentle introduction to k-fold cross-validation. https://machinelearningmastery.com/k-fold-cross-validation.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, page 160–172.

Carter, C. and Catlett, J. (1987). Assessing credit card applications using machine learning. *IEEE Expert*, 2:71–79.

Gandhi, R. (2018). Build your own convolution neural network in 5 mins.

Giannotti, F., Gozzi, C., and Manco, G. (2002). Clustering transactional data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 175–187. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Holm, M. (2018). Machine learning and spending patterns : A study on the possibility of identifying riskily spending behaviour. Master's thesis, KTH, School of Computer Science and Communication (CSC).

Khandani, A. E., Kim, A. J., and Lo, A. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking Finance*, 34(11):2767–2787.

Koehler, M., Tivnan, B., and Bloedorn, E. (2005). Generating Fraud: Agent Based Financial Network Modeling. page 5.

LeCun, Y. and Bengio, Y. (1995). *Convolutional networks for images, speech, and time-series*. MIT Press.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nat.*, 521(7553):436–444.

Lloyds Banking Group. Lloyds bank. 2022. who we are. https://www.lloydsbank.com/banking-with-us/who-we-are.html. Accessed: 2022-05-24.

Lopez-Rojas, E. and Axelsson, S. (2012). Multi agent based simulation (mabs) of financial transactions for anti money laundering (aml).

Lv, F., Huang, J., Wang, W., Wei, Y., Sun, Y., and Wang, B. (2019). A two-route cnn model for bank account classification with heterogeneous data. *PLOS ONE*, 14(8).

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Mirashk, H., Albadvi, A., Kargari, M., Javide, M., Eshghi, A., and Shahidi, G. (2019). *Using RNN to Predict Customer Behavior in High Volume Transactional Data*, pages 394–405.

Poggio, T., Lo, A. W., LeBaron, B. D., and Chan, N. T. (2001). Agent-based models of financial markets: A comparison with experimental markets. *SSRN Electronic Journal*.

Rekik, Y. M., Hachicha, W., and Boujelbene, Y. (2014). Agent-based modeling and investors' behavior explanation of asset price dynamics on artificial financial markets. *Procedia Economics and Finance*, 13:30–46.

Treanor, J. (2018). Don't bank on it: Why we fail to switch our accounts. https://www.bbc.co.uk/news/business-44522630.

van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Wang, L., Ahn, K., Kim, C., and Ha, C. (2018). Agent-based models in financial market studies. *Journal of Physics: Conference Series*, 1039:012022.

Zakrzewska, D. and Murlewski, J. (2005). Clustering algorithms for bank customer segmentation. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 197–202.