Bank Statement Transactions – Segmentation

By John McCabe

Link to project: Bank Statement Transactions Segmentation

Analysis Domain

The availability of bank statement data via companies such as Yodlee has resulted in a boom in FinTec companies. Products derived from this data include, financial management, wealth management, fraud prevention and accounting. The company I worked, developed a product using this data. The premise was that companies could benchmark their sales against competitors. The product was basic and took the data feeds from a data provider who had enriched each transaction with a merchant and category, then aggregated and published via Tableau.

The product failed, due to several factors, including:

- Sample bias, the profile was heavily skewed to young people
- Poor quality of merchant tagging, large number of untagged transactions
- No added value in the form of insight or segmentation of the sample; it is this failing that my project looks to address

The data provider removed all PII so the data has no personally attributable information in it.

Analysis Task

Aim: Enrich bank transaction data with additional attributes on the users to enable companies to identify their customer base with an aim to develop more effective marketing strategies.

Question: Is it possible to use bank transaction data to cluster the users into discreet segments?

Objectives: Using a sample of users**,** identify clusters with similar attributes. Apply as training data to form the basis of a model which can be pushed back to the full data set

Analysis Plan

Data Description

The data provider delivered three sets of data all linked via hashed id's, with a hierarchy of user-account-transactions. The transaction data included both deposits (positive) and spend (negative). Along with some basic demographics, age and gender the user data also had regional details and an associated mosaic group code which had been attached via the postcode so the accuracy is directional at a user level.

Users	Accounts	Transactions
Gender	Account Type	TransactionDate
YearofBirth		TransactionTag
RegionSID		TagName
PostcodeSector		Amount
MosaicGroup		

Planned Steps

Sampling

Due to volumes involved, over 100 million transactions per year, I intend to use stratified sampling, by region. In order to reduce any regional affects.

Data investigations

Split deposits and spend, analyse separately to simplify the investigation of the transactions. Due to the sampling, it is important to get metrics that scale out to the full sample. The aim will be to determine which features to use, the impact of outliers, the distributions of the features and whether any transformations are required.

Feature engineering

Review potential for reducing amount of untagged transactions for either the deposits or spend.

Flatten & combine

As all the data was stored in MSSQL, it is fully normalized. It will be necessary to extract, subsequently flatten and combine up to the user level.

Dimension reduction

As there are hundreds of tag names and therefore features it will be necessary to reduce the number of features otherwise the analysis will suffer the curse of dimensionality.

Clustering & review

Apply a clustering algorithm, to the resulting principal components. Using visualisations to establish the standout profiles of each cluster and determine a descriptive segment.

Update & model

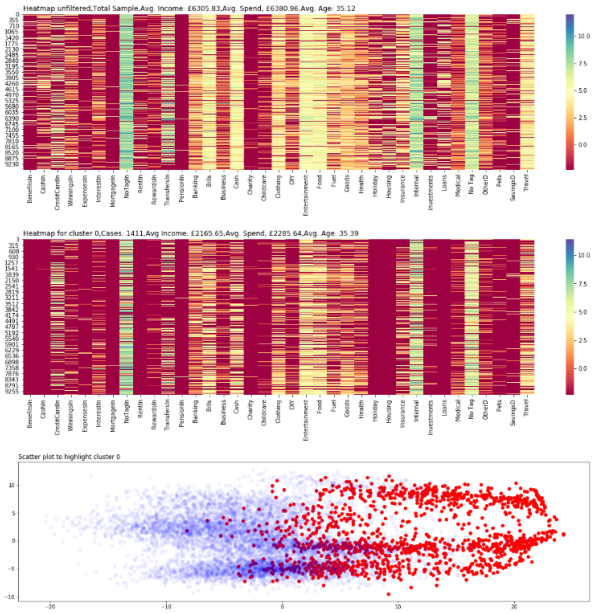
Use the resulting data set to model out to the full data set.

Findings

After executing the above plan, 5 clusters were identified, with varying coherency. Here I cover the main characteristics.

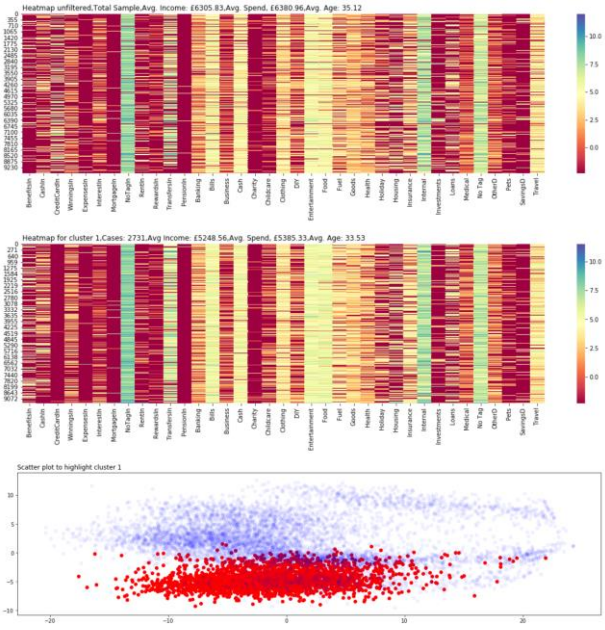
Using filtered heatmaps, highlighting which categories varied from the average for all users I was able to create descriptions (and glibly added names) for 5 clusters, some of the clusters. Below are the plots and descriptions of the clusters identified.

Cluster 0: Low income living with parents



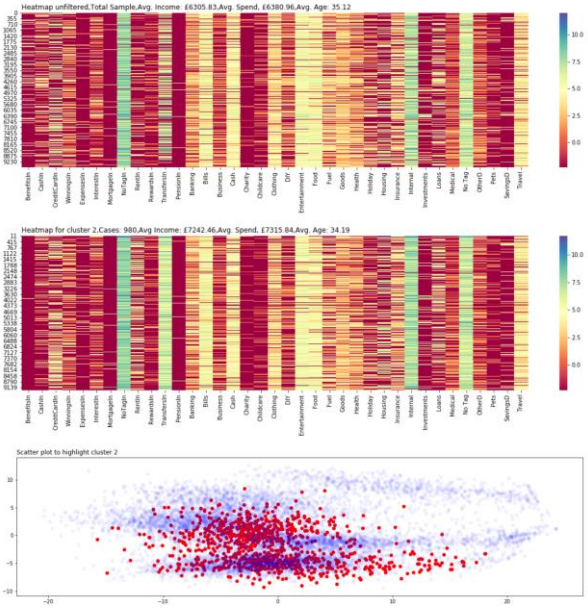
It is clear from the heatmaps and confirmed by the average income and spend that this cluster has low deposits and not significant uplift anywhere other than credit card. The minimal spend on housing and holiday could indicate that these users potentially could still live at home with their parents, though the average age is quite high for this to be the case. It appears there are 2 or 3 sub-clusters, which could be investigated but the number of cases is very low and there is a definite risk of over-fitting.

Cluster 1: Average income, families



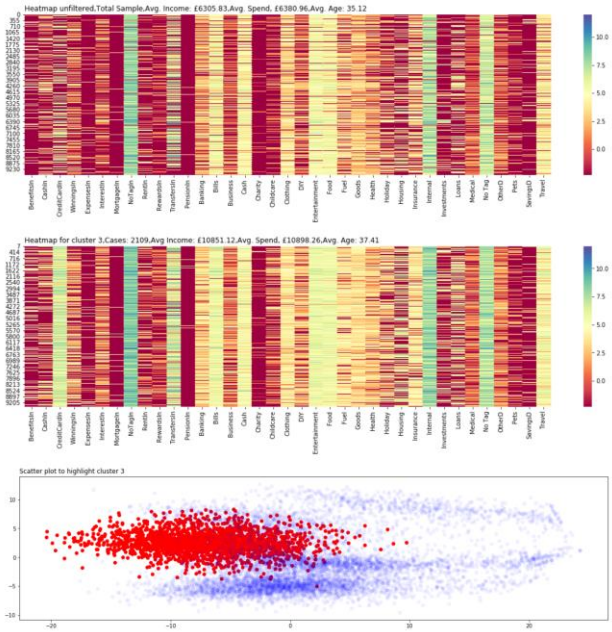
This is a very distinct segment with only partial overlaps with other data points. It presents a reasonably coherent story, the categories with lift, are categories assocaited with families and in particular mid-low income families.
One interesting aspect is the absence of credit card payments as this is something that I would associate more with this category, so it may not be as clear cut as it looks.

Cluster 2: Young people who love to transfer



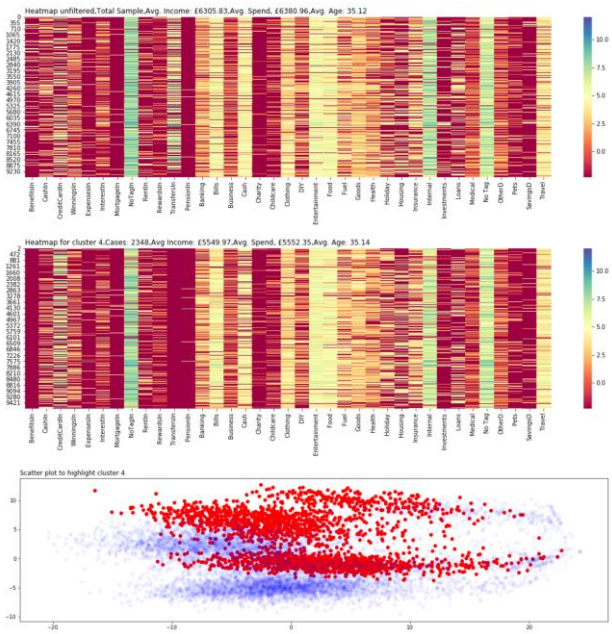
This is a slightly confused cluster with low sample numbers. Reviewing the scatterplot, it is clear that it spans a wide area and overlaps with a number of other clusters. I would deduce that this cluster holds a high proportion of young single users based on the markers highlighted.

Cluster 3: Well-heeled families



This is quite a compact cluster and as such I would expect it to be quite coherent. It is clear from the category markers and confirmed by the average spend that this cluster includes households that are about national average. Taking age into account, I would conclude that this cluster contained families in advanced stages of life a lot with children, hence the high spend on food and goods etc. The high income would point towards double incomes so I would expect there to be multiple incomes.

Cluster 3: Confused



The only factor that seems to drive this cluster is the lack of transfers into the account. The rest of the heatmap is broadly in line with the average. Reviewing the scatter plot, it would appear that there are at least 2 distinct clusters within this cluster, which is the likely cause of the confused heatmaps. Isolating and re-clustering may provide a clearer picture of the subsets of data within here.

One of the most interesting findings was, certainly for the more coherent clusters, the alignment with the Mosaic Groupings. This can be seen clearly in the notebook. Intentionally I ran the heatmaps and wrote the descriptions before adding in both the derived salary information and the Mosaic Grouping bar charts. Adding these showed consistencies with the evidence from the heatmaps. The derived salary information also showed consistency with the descriptions.

Reflections

Ultimately it was possible to identify some clusters using a clustering algorithm applied to bank transactions.

With only five clusters identified and some of them quite confused, it means that it is a long way from being operational but there was enough evidence to suggest that this is a viable approach.

The data is extremely rich, with the behaviours and transactions of users varying greatly, this became clear when reviewing the outliers. With more time I would have spent longer working through the outliers as I think it would have revealed, not just valid outliers but the way people use multiple accounts; transferring money between them. This would enable deduplication of deposit and spend, resulting in a clear picture of income and expenditure – this would be a very important feature.

There were a few areas where I was left deeply unsatisfied, they fall in two categories, the first is the quality of the data. There were several tagging issues that would have impacted the results:

a) The high volume of untagged transactions, with over 50% deposit and 20% of spend. This is a considerable amount of unaccounted transactions.
b) Mis-tagging of Tag Names. Several transactions were miscoded, this became apparent when reviewing the data when trying to group deposits and spend. The impact is magnified when grouping since one category is inflated while another is deflated.

The second is my lack of knowledge and expertise. The areas where I think my lack of experience/ knowledge impacted the results were:

a) The identification of salaries; my function for identifying salaries failed to pick up about 50% of users. When investigating it seemed that using Fourier analysis would be a better approach, but I was unable to get usable results. The identification of frequent transactions would be effective beyond salary information, it could isolate regular spend on food and bills; indeed, all transactions that are periodic, again this would deliver significant attributes.
b) Transformation of zero-inflated data. This I felt may have been the biggest single factor in not getting clearer clusters. I made the decision to go with log-transformation but in hindsight I think more investigations into ZIP and other strategies of handling zero-inflated data may have delivered better results. Although not included in the final notebook, I ran experiments with different scaling functions and none appeared to be as effective as the log-transformation, hence it was the one used.

Finally there was one area I would have liked to have explored in greater detail, the clustering algorithms. K-means didn't appear to be well suited to the data but I felt that both DBSCAN and GMM were both quite well suited to the data and in the end I went with GMM. In hindsight I am left wondering if DBSCAN would have been more effective. If the initial dense clusters had been removed and then re-run. Maybe even continue this process until all clusters had been identified, something on the to do list!

Paradoxically, I was left both enthused and disappointed with the project.