Receipt Bank Placement Report

Name: Joe Davies

Email: j.m.m.davies@qmul.ac.uk

Job Title: Data Scientist

Host Company: Receipt Bank

Logo:

Address of Host Company:

1st floor, 99 Clifton St, Hackney, London EC2A 4LG

Manager: Cameron Stone

Placement Officer: Dr Eram Rizvi

Start and End Date: 01/06/2019 - 25/09/2019

Introduction

Receipt Bank was started in 2010 and operates on a Software as a Service model. This means the service they provide is one you subscribe to and make monthly payments to keep. They aim to assist small/medium sized businesses by streamlining their accounting and bookkeeping processes. A business owner is able to take a picture of a receipt which is then put through a machine learning algorithm that extracts relevant data from the image and sends it off to their accountant/bookkeeper. The type of fields that are collected include: total amount spent, tax, supplier, date etc. On average this saves an accountant one hour per week per client in manual data entry and ordering of receipts. One accountant can have scores of clients so this software can save hundreds of hours per week of work.

The company works with around 300,000 businesses and processes 5 million receipts every month. The business has approximately 450 employees with offices in: the UK, France, USA, Australia and Bulgaria (where the bulk of the technological aspects are based).

This placement was carried out as part of the DISCnet-studentship funded PhD. I chose this particular placement because a large part of it was research and development using machine learning, something similar to the subject of my PhD. I also wanted a stepping stone between academia and the commercial world and felt this was a great example of that.

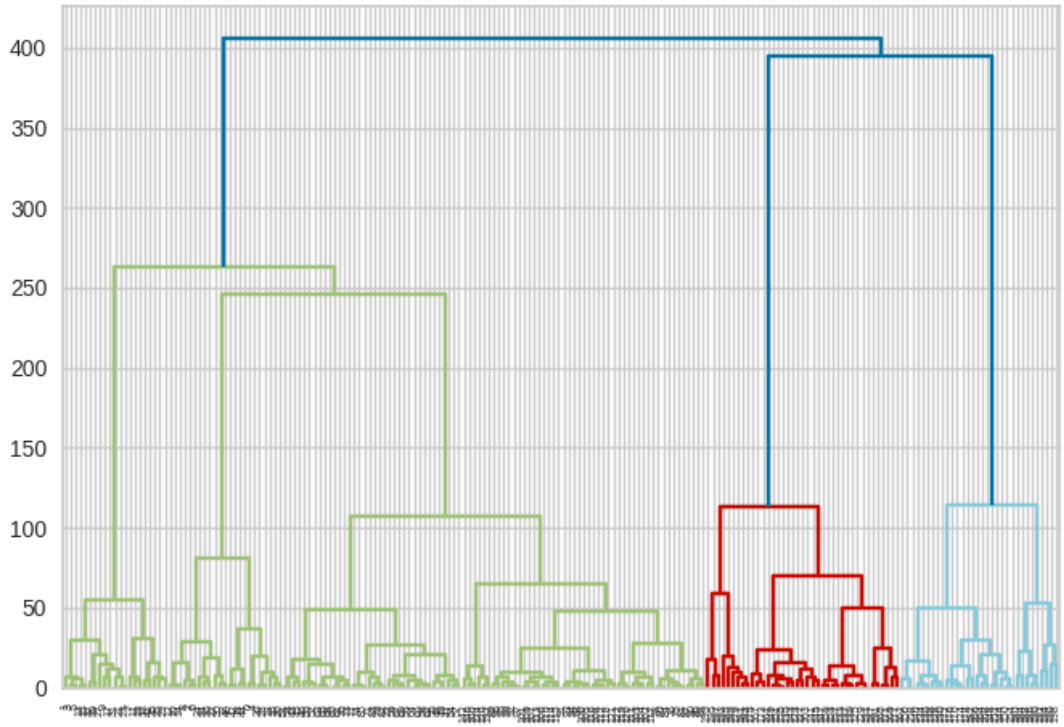
What did I do?

I was tasked with performing exploratory data analysis to segment the user base of Receipt Bank. This means trying to find clusters of users that interact with the service and seeing how they differ from each other. In doing this, one can find ways to target those users with more specialized services or even attract new users with something they may specifically be interested in. An example of this already in use is targeted advertisements where your past choices online influence how advertisers will try and market to you. If you buy a lot of shoes, they will offer you shoe related products and so on.

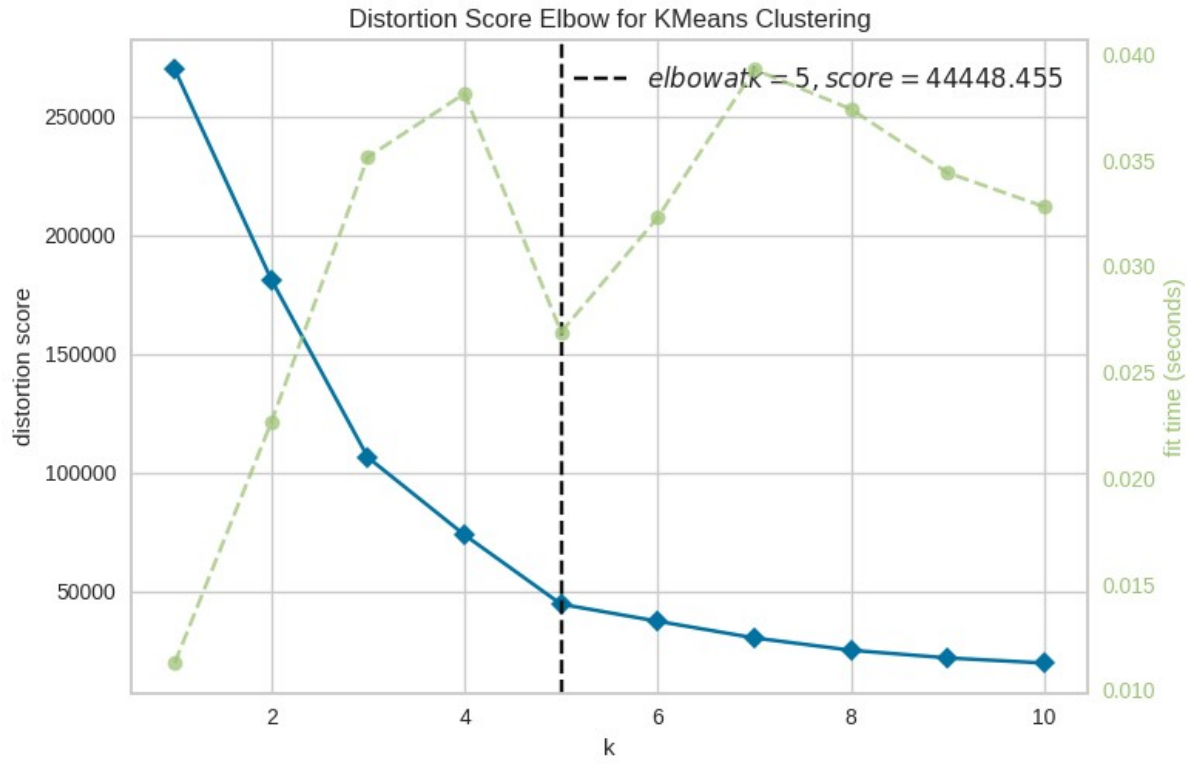
One of the techniques for doing this is clustering, a type of unsupervised machine learning that aims to find patterns in data that would be incredibly difficult to find for a human. Clustering aims to find those data points that have the shortest distance between them and then group those together to form cohorts of points which can be analyzed and understood. One of the most common algorithms for doing this is k-means clustering which partitions observations into k number of clusters. This was the machine learning aspect of my placement and was performed using the scipy and sklearn modules in python. An example of this can be seen below.

Here we can see five clusters each describing different types of people. For example, those with a high income and a high spending score are able to be marketed higher ticket items whereas those with a low income but high spending score might be better off being marketed some lower price options.

Other statistical tests were carried out in relation to this. One of the main things that needed to be done was finding how many clusters would be optimal to perform k-means with and there are a couple of ways I went about doing this. One of them was by doing hierarchical clustering which look to group points together and then group those groups and so on until there are just two groups. This produces something called a dendrogram which can be seen below, and is programmed using the scipy module in python.



We can use this to approximate number of clusters needed by drawing a line across where the bigger sub-clusters are and seeing how many lines we intersect. Another method is the elbow plot which uses k-means itself to find where the error rate begins to smooth out. An example where the correct number of clusters is five is shown below.



Other statistical tests performed included: Pearson correlation test to investigate whether there was a statistically significant relationship between variables, Principal Component Analysis which aims to lower the dimensionality of your dataset in order to make visualization easier and T-tests to see whether a subset of data is still a part of the full dataset.

The data that I was using was mostly taken from an SQL database using software called Looker. This meant that I had to learn some SQL but thankfully python has a module that allows one to load SQL data directly into the compiler you are using. I worked mainly with ‘event data’ or data that comes from when an event is triggered such as pressing a button on a website or receiving a notification. This was then paired with the user account that triggered the event and information on where that event made a change. Some accounts are able to change details of others and this is useful for an accountant that has clients also using the website/mobile app. The data was stored in a data warehouse which in turn drew data from a datalake. This meant it had to be cleaned for use and put into a form that I could manipulate, which mostly meant using a python module called pandas. Visualization was done using matplotlib and seaborn.

This exploratory data anlaysis was done using Jupyter notebooks. This facilitated me being able to report back to people as I was able to share my findings in a way that was easy to understand as visualizations could be put right next to the code that produced them.

Key Results

Over the course of the placement I managed to report back some of my findings. My analysis kickstarted an inter-departmental discussion about the problem I was trying to solve and through this we found that other teams in the company were doing the same thing from different viewpoints. I was doing anaysis based on no prior assumptions on the data in order to not produce any bias. Other teams relied on certain biases to make inferences and so were looking at specific users and case studies etc. This resulted in the formation of a group of people with different business backgrounds working together to tackle the problem from different angles. I was able to provide this group statistical reuslts that they could then work from. For example, I was able to tell them when people uploaded receipts most often in the different regions the Receipt Bank operates in and this let the tech team know when high traffic times would be. I was able to let the business analysts know how often clients submitted items and whether there was any relationship between event types and who fired them.

Another key results was finding out which fields were edited most often on receipts. This is a crucial one in lowering costs for the company. After the machine learning algorithm is applied to a receipt, it is then sent off to have any fields that didn't reach an accuray threshold to a physical person to check, which costs money. I was able to narrow down which of these fields are more common in order that we only pay to get those particular fields checked.

Differences between Academia and Commercial

Going from the academic to commercial spaces was definitely a challenge. One of the main differences I found was the interaction with the data. In academia, when the data gets to me it has been cleaned and sorted in a way that makes it fairly easy to manipulate, with 'best practice' documents showing how data should be stored and in what format. In the commercial realm this is not the case, with much of the data being unstructured to begin with and requiring you to clean and order it. It's a running joke, for example, that every data scientist has a way of converting one form of the data into another, as this is often a point of contention. Most of my time was spent cleaning and organizing the data into a form that could be used in the various algorithms I was running.

Another difference are the permissions. In business every decision must comply with strict GDPR guidelines and software cannot be downloaded without the prior consent of your tech lead or IT support officer. This makes sense here as user data is what is at risk, something that could feasibly be used to identify someone. This can lead to some issues though, mainly the time it takes to get this software is increased and, in my case, sharing software can be a problem. The notebook that I often used to code with was on a shared server and so if someone else wanted to use it and my PC was idle because I was in a meeting/at lunch etc I would have to wait to use it again. This isn't something experienced as much in academia, as we generally don't work with user data. Not only this, but we have fewer entries of attack for anyone that would like to take our data because we don't have a website/app that 'users' interact with in the same way that Receipt Bank does.

There are also differences in how the two types interact with each other. In academia there is more of an emphasis on formality with conferences and meetings being as concise as they can be. There are fewer outings for an outings sake compared to commercial, which is a much more social

workspace. I was only around for 3 months but in that time I attended many gatherings where the only reason was to increase team work and team cohesion, which I enjoyed. This is a larger part of the commercial arena as teams must work together well in order to get good results. This can provide a higher pressure environment than academia as well, necessitating outings.

Conclusion

Overall I really enjoyed my placement. The ability to learn techniques like data manipulation, cleaning and wrangling will no doubt help me when it comes to the programming side of my PhD. Specifically, the techniques on dimensionality reduction, data frame handling and statistical tests will be useful. Also, practicng presentations and keeping an audience interested in analyses will be invaluable and I can further see the business acumen gained being useful if I leave academia.

I also learnt some good ways of having a healthy work-life balance. Not stressing and fretting over what is happening at work is key and communicating any worries with collegues was a real help. I am thankful to Receipt Bank and especially my manager Cameron for taking my under their wing!