**Aksheyen and Georgios**

**Santander Recommendation Project**

Aim: The project serves as an introduction to machine learning, to develop our existing skills in python

Tech Stack:

* Python: The libraries we used are numpy, pandas, graphlab and sklearn. Pandas were used because we needed to put the available data into a pandas DataFrame. Numpy was also helpful when we had to create an array with the users ( or ncodpers).
* GraphLab: A high performance distributed computation framework which is heavily used in industry for machine learning tasks. This is the reason we chose it, aiming to build an item recommendation engine after we brought the data in the appropriate format to be used by Graphlab.

Method:

1. Downloaded data in CSV format and opened using python in a Dataframe as in figure1.

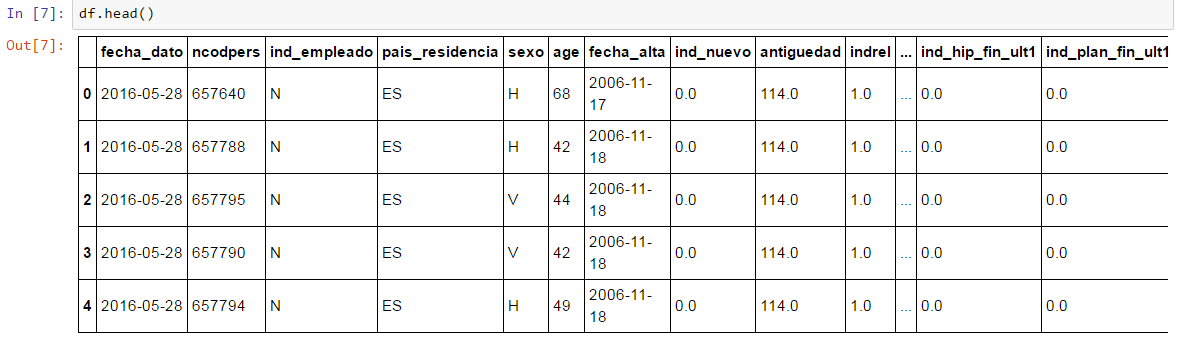


Figure 1 Dirty data in pandas DataFrame

1. The data was cleaned and explored to find most popular products and the most important features for classifying customers as can be seen in the two figures below.

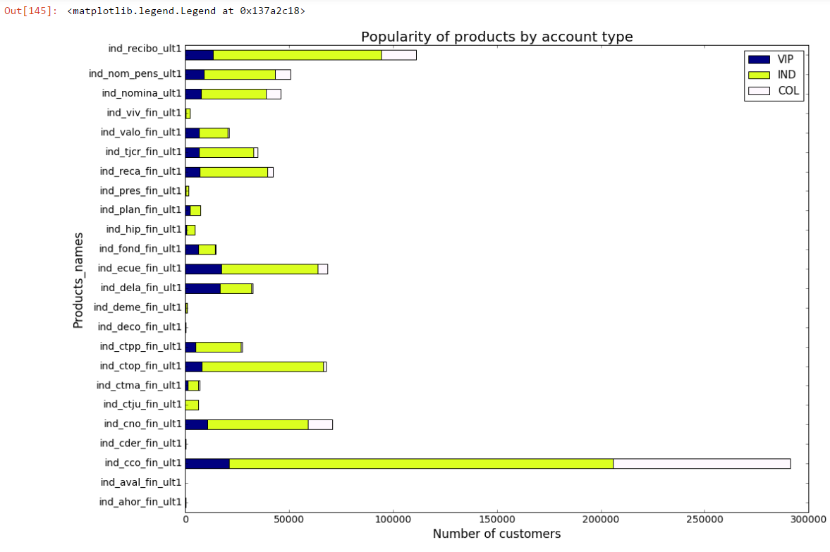
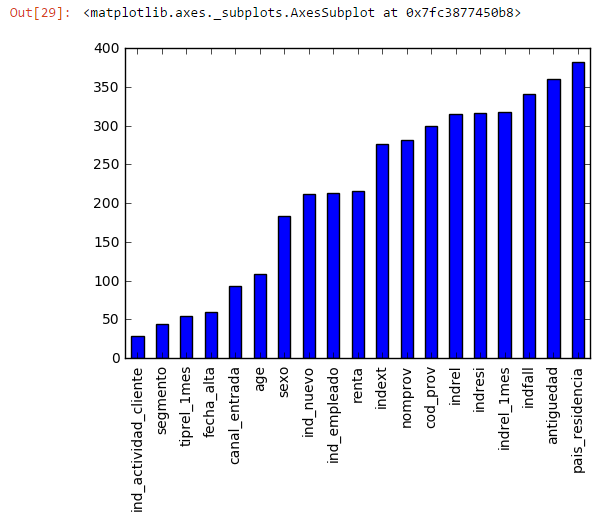


Figure 2 Computing the most important features for classifying clients

Figure 3 Data exploration. The most popular products by account types VIP, Individual and College

1. The data was engineered into the format required by GraphLab, which is in essence a matrix with all the products a user has or hasn’t bought.

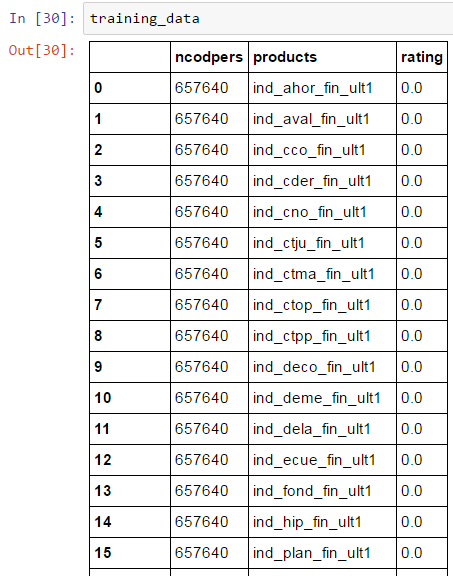
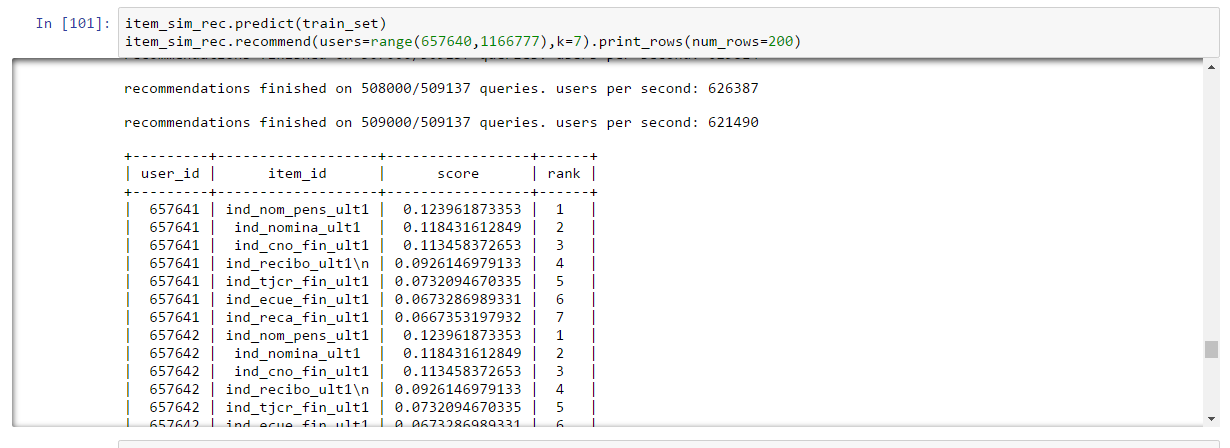
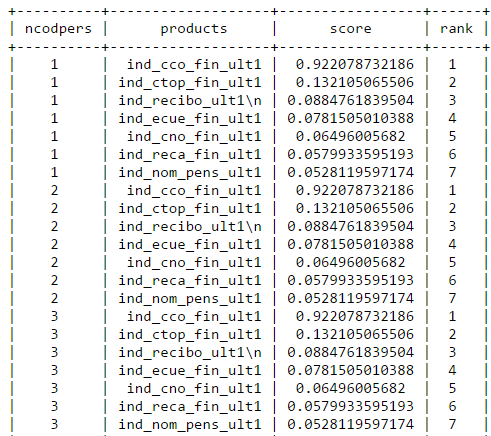
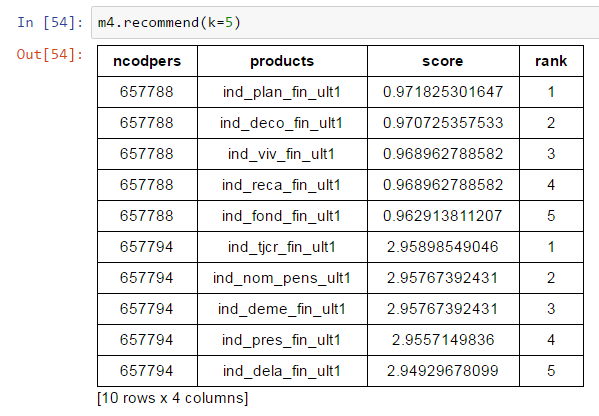


Figure 4 The format of data required for GraphLab

1. Item-Item Similarity recommendation engine – Jaccard



1. Ranking factorization Similarity Model
2. We noticed the products recommended were not specifically tailored to individual customers. This may have been because of the binary format of the ratings of the products, the large dataset and the number of 0s. To test the algorithms we built a recommendation engine on a much smaller dataset of 55 rows, with randomized ratings figures ranging from 1-5 using item-item similarity recommendation. This resulted is more personalized recommendations.
3. This concludes the algorithms on GraphLab are not very effective with binary datasets.